电子工程应用精讲系列

SPSS 17.0中文版常用功能与应用实例精讲

赖国毅 陈超 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是畅销书《SPSS 15.0 中文版常用功能与应用实例精讲》的升级版。全书从实用的角度出发，通过常用模块与大量工程实例相结合的形式，深入浅出地介绍了 SPSS 17.0 中文版的操作方法和应用技巧。全书分为三篇共 20 章，第 1～4 章为基础知识篇，简单介绍了 SPSS 软件用户界面、参数设置、数据文件的基本操作、简单数据处理和分析结果表示，引导读者入门。第 5～15 章为 SPSS 统计分析常用模块篇，介绍了均值与方差分析、参数检验、非参数检验、相关分析、回归分析、聚类分析、判别分析、因子分析与主成分分析、可靠性分析、时间序列分析和生存分析等模块，并结合实例重点解释了各个模块的算法原理和基础应用。第 16～20 章为行业应用实例篇，详细介绍了 SPSS 在实际工程领域中的应用，具体包括：社会调查与统计、市场研究/企业数据分析、证券金融统计、医学统计和生物学统计，实例典型、代表性和指导性强。读者通过学习，可以举一反三，进一步加深巩固，实现从入门到精通。

本书适合 SPSS 的初、中级读者使用，是统计学、社会学、经济学等专业的理想教材，也是读者进行课题研究及定量分析的首选参考书。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

SPSS 17.0 中文版常用功能与应用实例精讲／赖国毅，陈超编著. —北京：电子工业出版社，2010.8
（电子工程应用精讲系列）
ISBN 978-7-121-11330-7

Ⅰ. ①S… Ⅱ. ①赖… ②陈… Ⅲ. ①统计分析－软件包，SPSS 17.0 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2010）第 132271 号

责任编辑：高洪霞
印　　刷：北京东光印刷厂
装　　订：三河市皇庄路通装订厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×1092　1/16　印张：34.5　字数：758 千字
印　　次：2010 年 8 月第 1 次印刷
印　　数：4000 册　定价：62.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言



SPSS 软件是一款优秀的统计分析工具，在调查统计行业、市场研究行业、医学统计、政府和企业的数据分析中应用非常广泛，同时也是世界上最早的统计分析软件之一。纵观现已出版的关于 SPSS 的同类书中，大多是基于 SPSS 15.0 之前的英文界面下的老版本，同时由于实用性不强，缺少工程实例导航，这给国内的 SPSS 初学者带来很多困难。本书的出版旨在弥补这种不足。

本书内容

本书以全新的 SPSS 17.0 中文版为写作平台，从应用和实用的角度出发，通过常用功能与大量实例相结合的形式，深入浅出地介绍了 SPSS 17.0 中文版的操作方法和应用技巧。本书分为三篇共 20 章，具体内容如下。

基础知识篇

包括第 1～4 章。简单介绍了 SPSS 软件用户界面、参数设置、软件基本操作、统计图绘制和基础统计描述，引导读者进行 SPSS 入门。

- 第 1 章　SPSS for Windows 简介
- 第 2 章　数据的基本操作
- 第 3 章　统计图表绘制
- 第 4 章　基础统计描述

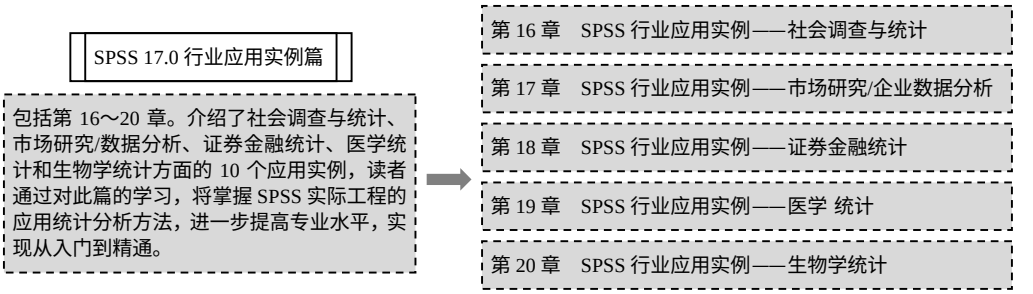
SPSS 统计分析常用模块篇

包括第 5～15 章。介绍了均值与方差分析、参数检验、非参数检验、相关分析、回归分析、聚类分析、判别分析、因子分析与主成分分析、可靠性分析、时间序列分析和生存分析等模块，在解释各个模块算法原理的同时，结合多个实例进行说明，方便读者理解和巩固。

- 第 5 章　均值分析与方差分析
- 第 6 章　参数检验
- 第 7 章　非参数检验
- 第 8 章　相关分析
- 第 9 章　回归分析
- 第 10 章　聚类分析
- 第 11 章　判别分析
- 第 12 章　因子分析与主成分分析
- 第 13 章　可靠性分析
- 第 14 章　时间序列分析
- 第 15 章　生存分析



前　言



第三篇为行业应用实例篇，包括第 16～20 章。介绍了社会调查与统计、市场研究/数据分析、证券金融统计、医学统计和生物学统计方面的 10 个应用实例，读者通过对此篇的学习，将掌握 SPSS 实际工程的应用统计分析方法，进一步提高专业水平，实现从入门到精通。

本书特色

与同类书相比，本书主要特色如下。

01 基于全新的 SPSS 17.0 中文版环境，内容从零开始、由简到难，常用操作与大量实例相结合，边讲边练。非常利于初学者和英语不熟的读者上手。

02 本书提供了 11 个 SPSS 常用模块设计实例，这些实例实用、基础、典型，易学易懂，几乎囊括了 SPSS 的所有内容，为读者提供了良好的学习途径。

03 通过 10 个行业工程实例对 SPSS 软件进行了综合性应用，范围涉及社会调查与统计、市场研究/数据分析、证券金融统计、医学统计和生物学统计领域，便于读者举一反三，使应用水平产生质的飞跃。

适合读者群

本书适合 SPSS 的初、中级读者使用，同时也是广大读者进行统计学、社会学、经济学等领域课题研究及定量分析的必备参考宝典。

本书作者

本书由赖国毅、陈超主持编写，参加本书编写工作的还有：赵汶、唐清善、邱宝良、李宁宇、严剑忠、黄小宽、付军鹏、张广安、贾素龙、金平、徐春林、谢正义、郑贞平。

由于作者的水平有限，书中难免存在一些不足之处，敬请广大读者批评和指正。欢迎读者通过电子邮箱 jsj@phei.com.cn 与我们联系。

目　录

第一篇　基础知识篇

第 1 章　SPSS for Windows 简介

1.1　SPSS for Windows 概述 …… 2
1.2　SPSS 的运行方式 …… 3
1.3　SPSS 模块与安装 …… 3
　1.3.1　SPSS for Windows 17.0 模块介绍 …… 3
　1.3.2　SPSS for Windows 17.0 安装步骤 …… 4
1.4　SPSS 的主要界面 …… 9
　1.4.1　SPSS 的启动 …… 9
　1.4.2　SPSS 的数据编辑窗口 …… 10
　1.4.3　SPSS 的结果输出窗口 …… 14
1.5　本章小结 …… 17

第 2 章　数据的基本操作

2.1　建立数据文件 …… 18
　2.1.1　输入数据建立数据文件 …… 19
　2.1.2　直接打开其他格式的数据文件 …… 19
　2.1.3　使用数据库查询建立数据文件 …… 20
　2.1.4　导入文本文件建立数据文件 …… 22
2.2　编辑数据文件 …… 22
　2.2.1　输入数据 …… 23
　2.2.2　定义数据属性 …… 23
　2.2.3　插入或删除数据 …… 32
　2.2.4　数据的排序 …… 33
　2.2.5　选择个案 …… 34
　2.2.6　转置数据 …… 36
　2.2.7　合并数据文件 …… 37
　2.2.8　数据的分类汇总 …… 42
　2.2.9　数据菜单的其他功能 …… 45
2.3　数据加工 …… 46
　2.3.1　数据转换 …… 46
　2.3.2　数据手动分组（编码） …… 49
　2.3.3　数据自动分组（编码） …… 53
　2.3.4　产生计数变量 …… 54
　2.3.5　数据秩（序）的确定 …… 56
　2.3.6　替换缺失值 …… 59
2.4　数据文件的保存或导出 …… 60
　2.4.1　保存数据文件 …… 60
　2.4.2　导出数据文件 …… 61
2.5　本章小结 …… 62

第 3 章　统计图表绘制

3.1　条形图 …… 63
3.2　线图 …… 69
3.3　面积图 …… 71
3.4　饼图 …… 73
3.5　高低图 …… 75
3.6　箱图 …… 77
3.7　直方图 …… 79
3.8　本章小结 …… 80



第 4 章　基础统计描述

4.1　数理统计量概述 …… 81
4.1.1　均值（Mean）和均值标准误差（S.E. mean） …… 81
4.1.2　中位数（Median） …… 82
4.1.3　众数（Mode） …… 82
4.1.4　全距（Range） …… 83
4.1.5　方差（Variance）和标准差（Standard Deviation） …… 83
4.1.6　峰度（Kurtosis）和偏度（Skewness） …… 84
4.1.7　四分位数（Quartiles）、十分位数（Deciles）和百分位数（Percentiles） …… 85
4.2　数据描述 …… 85
4.3　频数分析 …… 88
4.4　探索分析 …… 92
4.5　交叉列联表分析 …… 97
4.6　比率分析 …… 103
4.7　P-P 图、Q-Q 图 …… 106
4.8　本章小结 …… 108

第二篇　SPSS 统计分析常用模块篇

第 5 章　均值分析与方差分析

5.1　均值分析 …… 111
5.2　方差分析概述 …… 117
5.2.1　方差分析的基本原理 …… 117
5.2.2　方差分析的概念和假设 …… 119
5.3　单因素方差分析 …… 119
5.4　多因素方差分析 …… 129
5.5　重复度量 …… 139
5.5.1　数据重构 …… 141
5.5.2　重复度量 …… 143
5.6　协方差分析 …… 149
5.6.1　非饱和模型的 SPSS 操作 …… 150
5.6.2　饱和模型的 SPSS 操作 …… 153
5.7　本章小结 …… 155

第 6 章　参数检验

6.1　参数检验概述 …… 157
6.1.1　参数检验问题的提出 …… 157
6.1.2　参数检验的基本步骤 …… 158
6.2　单样本 T 检验 …… 159
6.2.1　单样本 T 检验的基本方法 …… 159
6.2.2　单样本 T 检验的实例 …… 160
6.3　独立样本 T 检验 …… 164
6.3.1　独立样本 T 检验的基本方法 …… 164
6.3.2　独立样本 T 检验的实例 …… 165
6.4　配对样本 T 检验 …… 168
6.4.1　配对样本 T 检验的基本方法 …… 168
6.4.2　配对样本 T 检验的实例 …… 169
6.5　本章小结 …… 172

第 7 章　非参数检验

7.1　卡方检验 …… 174
7.1.1　卡方检验的基本原理 …… 174

7.1.2 卡方检验的实例 …… 175
7.2 二项分布检验 …… 179
7.2.1 二项分布检验的基本原理 …… 179
7.2.2 二项分布检验的实例 …… 179
7.3 游程检验 …… 182
7.3.1 游程检验的基本原理 …… 182
7.3.2 游程检验的实例 …… 182
7.4 单个样本 K-S 检验 …… 185
7.4.1 单个样本 K-S 检验的基本原理 …… 185
7.4.2 单个样本 K-S 检验的实例 …… 186
7.5 两个独立样本的非参数检验 …… 190
7.5.1 两个独立样本 Mann-Whitney U 检验 …… 191
7.5.2 两个独立样本非参数检验的实例 …… 191
7.6 K 个独立样本的非参数检验 …… 195
7.7 两个相关样本的非参数检验 …… 200
7.8 K 个相关样本的非参数检验 …… 204
7.9 本章小结 …… 207

第 8 章 相关分析

8.1 简单相关分析 …… 208
8.1.1 简单相关分析的基本思想 …… 208
8.1.2 相关系数 …… 209
8.1.3 简单相关分析的操作步骤 …… 211
8.2 散点图 …… 213
8.3 偏相关分析 …… 216
8.4 本章小结 …… 218

第 9 章 回归分析

9.1 线性回归 …… 220
9.1.1 线性回归的基本原理 …… 220
9.1.2 SPSS 线性回归 …… 223
9.1.3 一元线性回归 …… 229
9.1.4 多元线性回归 …… 236
9.2 非线性回归 …… 241
9.2.1 非线性回归的基本原理 …… 241
9.2.2 SPSS 非线性回归 …… 241
9.3 曲线估计 …… 248
9.4 logistic 回归 …… 254
9.5 本章小结 …… 263

第 10 章 聚类分析

10.1 聚类分析概述 …… 265
10.1.1 聚类分析的基本原理 …… 265
10.1.2 聚类分析的基本方法 …… 266
10.1.3 不相似测度的度量方法 …… 268
10.2 系统聚类分析 …… 271
10.3 二阶聚类分析 …… 280
10.4 K-均值聚类分析 …… 292
10.5 本章小结 …… 300

第 11 章 判别分析

11.1 判别分析的基本原理 …… 303
11.2 SPSS 判别分析 …… 303
11.3 本章小结 …… 314



第 12 章　因子分析与主成分分析

12.1　基本原理 …… 316
12.1.1　主成分分析 …… 316
12.1.2　因子分析 …… 316
12.1.3　方法用途 …… 317
12.1.4　使用条件 …… 317
12.1.5　常用概念和分析步骤 …… 318
12.1.6　主成分和公因子数量的确定 …… 318
12.1.7　主成分和因子分析的联系与区别 …… 319
12.2　因子分析 …… 320
12.2.1　因子分析的数学模型 …… 320
12.2.2　因子分析的基本方法 …… 320
12.2.3　因子分析的实例 …… 321
12.3　主成分分析 …… 327
12.3.1　主成分分析的数学模型 …… 328
12.3.2　主成分分析的基本方法 …… 328
12.3.3　主成分分析的实例 …… 329
12.4　本章小结 …… 338

第 13 章　可靠性分析

13.1　可靠性分析概述 …… 340
13.1.1　可靠性分析的基本原理 …… 340
13.1.2　可靠性分析的基本方法 …… 341
13.2　SPSS 可靠性分析 …… 342
13.3　可靠性分析的其他问题 …… 348
13.3.1　拆半可靠性系数模型 …… 348
13.3.2　Guttman 模型 …… 351
13.3.3　平行模型 …… 352
13.4　本章小结 …… 353

第 14 章　时间序列分析

14.1　时间序列分析概述 …… 355
14.1.1　时间序列数据及其分析方法 …… 355
14.1.2　时间序列分析的模型、公式和记号 …… 355
14.1.3　SPSS 时间序列分析功能 …… 359
14.2　时间序列数据的预处理 …… 359
14.2.1　定义日期变量 …… 360
14.2.2　创建时间序列 …… 361
14.2.3　填补缺失数据 …… 365
14.3　指数平滑法 …… 367
14.4　自回归模型 …… 371
14.5　ARIMA 模型 …… 378
14.6　周期性分解 …… 386
14.7　本章小结 …… 390

第 15 章　生存分析

15.1　生存分析概述 …… 392
15.2　寿命表法 …… 393
15.3　Kaplan-Meier 分析 …… 398
15.4　Cox 回归分析 …… 405
15.5　本章小结 …… 412

第三篇 SPSS 17.0 行业应用实例篇

第 16 章 SPSS 行业应用实例——社会调查与统计

16.1 全国各地区农民家庭收支的分布规律分析 …… 414
16.1.1 实例内容说明 …… 414
16.1.2 实现方法分析 …… 415
16.1.3 具体操作步骤 …… 416
16.2 判定新生婴儿的性别比例是否正常 …… 434
16.2.1 实例内容说明 …… 434
16.2.2 实现方法分析 …… 435
16.2.3 具体操作步骤 …… 436

第 17 章 SPSS 行业应用实例——市场研究/企业数据分析

17.1 某商品销售量与超市规模和摆放位置的方差分析 …… 440
17.1.1 实例内容说明 …… 440
17.1.2 实现方法分析 …… 441
17.1.3 具体操作步骤 …… 442
17.2 汽车配件公司某产品尺寸的参数检验分析 …… 455
17.2.1 实例内容说明 …… 455
17.2.2 实现方法分析 …… 455
17.2.3 具体操作步骤 …… 456

第 18 章 SPSS 行业应用实例——证券金融统计

18.1 股票日历效应分析 …… 461
18.1.1 实例内容说明 …… 461
18.1.2 实现方法分析 …… 462
18.1.3 具体操作步骤 …… 462
18.2 汇率影响因素分析 …… 489
18.2.1 实例内容说明 …… 489
18.2.2 实现方法分析 …… 490
18.2.3 具体操作步骤 …… 491

第 19 章 SPSS 行业应用实例——医学统计

19.1 两组不同治疗方法的生存率分析 …… 513
19.1.1 实例内容说明 …… 513
19.1.2 实现方法分析 …… 514
19.1.3 具体操作步骤 …… 514
19.2 判断不同群体的患病比例有无差异 …… 520
19.2.1 实例内容说明 …… 520
19.2.2 实现方法分析 …… 521
19.2.3 具体操作步骤 …… 521

第 20 章 SPSS 行业应用实例——生物学统计

20.1 判断某种元素的缺乏是否对另一种元素含量有影响 …… 530
20.1.1 实例内容说明 …… 530
20.1.2 实现方法分析 …… 531
20.1.3 具体操作步骤 …… 532
20.2 根据动物身体属性数据分析个体之间的相似性 …… 535
20.2.1 实例内容说明 …… 535
20.2.2 实现方法分析 …… 535
20.2.3 具体操作步骤 …… 535

第一篇

基础知识篇

- 第 1 章　SPSS for Windows 简介
- 第 2 章　数据的基本操作
- 第 3 章　统计图表绘制
- 第 4 章　基础统计描述

第1章 SPSS for Windows 简介

SPSS 是世界上应用最为广泛的统计分析软件之一，已经成为社会学家和科学工作者统计分析时必不可少的工具。本章将从简要介绍 SPSS 17.0 for Windows 中文版的界面、操作等内容，引导读者入门。

本章学习要点：

- SPSS 17.0 for Windows 中文版的安装；
- SPSS 数据编辑界面；
- SPSS 结果输出界面；
- SPSS 的启动和退出。

1.1 SPSS for Windows 概述

SPSS 是“Statistical Package for the Social Sciences”的英文缩写，即“社会科学统计软件包”。随着 SPSS 产品服务领域的扩大和服务程度的加深，SPSS 公司于 2000 年正式将英文全称更改为“Statistical Product and Service Solutions”，意为“统计产品与服务解决方案”。

SPSS 软件是一款在调查统计行业、市场研究行业、医学统计、政府和企业的数据分析应用中享有盛名的统计分析工具，是世界著名的统计分析软件。1968 年，美国斯坦福大学的三位专业背景截然不同的研究生 Norman H. Nie、C. Hadlai (Tex) Hull 和 Dale H. Bent 开发了最早的统计分析软件系统，并基于这一系统于 1975 年在芝加哥合伙成立了 SPSS 公司。迄今为止，全球约有 25 万家公司在使用 SPSS，它们分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，全球 500 强中 80%的公司都在使用 SPSS，而其在市场研究和市场调查领域更是超过了 80%的市场占有率。因此，SPSS 是世界上应用最为广泛的统计分析软件。

SPSS 的基本功能有数据管理、统计分析、图表分析、输出管理等，具体内容包括描述统计、列联表分析，总体的均值比较、相关分析、回归模型分析、聚类分析、主成分分析、时间序列分析、非参数检验等多个大类，每个类中还有多种专项的统计方法。例如，回归分析分为线性回归分析、曲线估计、Logistic 回归等统计过程，并且每个统计过程中都允许用户选择不同的方法及参数。同时，SPSS 还设有专门的绘图系统，可以根据使用者的需要将给出的数据绘制成各种图形，以满足用户的不同需求。

目前，SPSS 已具有适合 DOS、Windows、UNIX 等多种操作系统使用的产品，国内常用的是其适用于 DOS 和 Windows 的版本。SPSS for Windows 软件包可以运行在微软公司的 Windows 98、Windows NT 4.0、Windows Me、Windows 2000、Windows XP、Windows Vista 和 Windows 7 等多个操作系统之下。本书将以最新的 SPSS 17.0 中文版为环境平台，详细介绍 SPSS 提供的各项专业统计功能。另外，本书所有的 SPSS 统计功能都附有对应的英文词汇，方便读者尽快熟悉英文统计术语，也适用于使用 SPSS 17.0 英文版的用户。

1.2　SPSS 的运行方式

SPSS 提供了三种基本运行方式：完全窗口菜单运行方式、程序运行方式和批处理方式。

完全窗口菜单运行方式简单明了，除人工输入数据需要键盘外，大部分操作命令和统计分析方法的实现都是通过菜单、图标按钮、对话框来完成的，使用者无需掌握编程就可以使用 SPSS 软件，非常适合传统的统计分析人员。

程序运行方式和批处理方式则是从使用者特殊的分析需求出发，编写 SPSS 命令程序，通过语句直接运行的。这两种运行方式要求使用者掌握专业的 SPSS 编程语法，对使用者的要求较高。

1.3　SPSS 模块与安装

SPSS 17.0 中文版软件包可以工作在两种模式下：单机模式和网络系统的用户界面模式。由于统计分析软件的数据量比较大，所以系统运行需要大于 16MB 以上的内存空间。

1.3.1　SPSS for Windows 17.0 模块介绍

SPSS for Windows 17.0 中文版共有 14 个部分，包括：基础工具（SPSS Base）、高级工具（SPSS Advanced Models）、分类数据（SPSS Categories）、分类树（SPSS Classification Trees）、复杂抽样数据（SPSS Complex Samples）、联合分析（SPSS Conjoint）、数据检查（SPSS Data Validation）、精确检验（SPSS Exact Tests）、地图分析（SPSS Maps）、缺失值分析（SPSS Missing Value Analysis）、程序设定（SPSS Programmability）、回归模块（SPSS

Regression Models）、统计表格（SPSS Tables）和预测工具（SPSS Trends）。用户可根据自身需求进行选择性安装，这样既节省硬盘空间，又方便使用。

其中，基础工具（SPSS Base）和高级工具（SPSS Advanced Models）的主要统计功能如表 1.1 和表 1.2 所示。

表 1.1 “SPSS 基础工具”主要统计功能

SPSS Base（基础工具）	
ACF（时间序列研究中的自动相关分析）	Mconvert（矩阵转化）
Aggregate（数据文件的汇总）	Means（均数及均数差别的显著性检验）
Anova（方差分析）	Mult Response（多变量数据的处理）
Autorecode（变量自动赋值处理）	Nonpar Corr（非参数资料的相关分析）
Correlations（相关分析）	Npar Tests（非参数检验）
Crosstabs（列联表处理）	Oneway（单因素方差分析）
Curvefit（11 种曲线模型的拟合）	Partial Corr（偏相关分析）
Date（变量定义与数据录入）	Plot（曲线绘制）
Descriptives（均数、标准差等描述性统计及 Z-分数转换）	Rank（等级排序、计算正态分数、百分比等分析）
Examine（数值分布形式的探究）	Regression（回归分析）
Fit（定义程序运行条件）	Report（结果输出）
Flip（数据行列转换）	Sort（数据排序）
Frequencies（频数表分析）	SP Chart（高分辨率的统计制图）
Graph（统计图制作）	Sysfile Info（显示 SPSS 格式的系统文件信息）
List（原始数据显示）	TS Plot（时间序列资料的统计制图）
Matrix Data（数据的矩阵处理）	T-Test（t-检验）

表 1.2 “SPSS 高级工具”主要统计功能

SPSS Advanced Models（高级工具）	
Cox Regression（Cox 回归模型）	Manova（协方差分析）
Hiloglinear（多因子系统模式的对数线性模型）	Matrix（高级矩阵转换）
Kaplan-Meier（Kaplan-Meier 生存时间模型）	Nonlinear（非线性分析）
Loglinear（对数线性模型及最优化检验）	Probit（依照所需概率做拟合最优化分析）
Logistic（Logistic 模型）	Survival（寿命表方式的生存分析）

1.3.2 SPSS for Windows 17.0 安装步骤

01 启动 Windows，打开“我的电脑”窗口，找到安装盘所在的驱动器，双击打开。安装程序自动运行，出现如图 1.1 所示的安装选项对话框。SPSS for Windows 17.0 中文版的安装程序仍是英文界面，下面进行简单的说明。

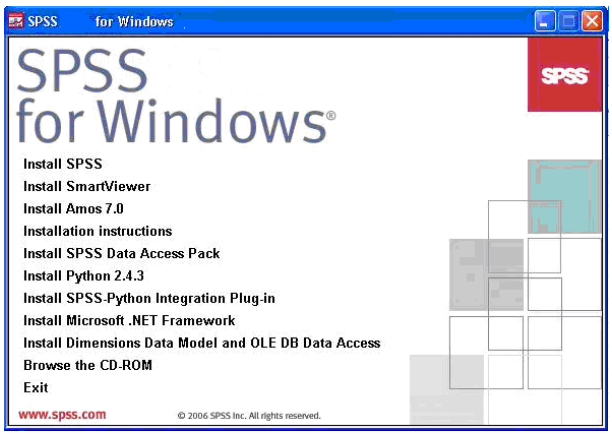


图 1.1　SPSS 17.0 for Windows 安装界面

单击安装界面上的选项，可以分别安装下面的模块。

- SPSS：SPSS 软件主体部分。
- SPSS SmartViewer：SPSS 提供的共享表格、图片的工具。
- Amos 7.0：结构方程建模软件。
- Instructions：SPSS 相关说明，包括软件安装说明、各部分软件介绍等。
- SPSS Data Access Pack：数据库挖掘软件，可与大多数主流数据库直接连接。
- Python 2.4.3：面向对象的 Script 程序语言编辑软件，可用于 SPSS 命令程序的编写。
- SPSS-Python Integration Plug-in：作为 SPSS 与 Python 之间的连接，在 SPSS 环境中，可在语法视窗撰写 Python 程序来执行 SPSS 一连串的分析动作。
- Microsoft .NET Framework：Windows 的托管代码编程模型，支持 SPSS 程序的编写。
- Dimensions Data Model and OLE DB Data Access：多维数据模型和一致数据访问软件。
- Browse the CD-ROM：浏览 CD 光盘，自主选择安装程序。
- Exit：退出安装程序。

本书主要针对 SPSS 主体部分程序的安装和统计功能的使用方法加以说明，其他软件的安装及使用，请查阅对应软件的相关书籍。

02 单击“Install SPSS（安装 SPSS）”命令，SPSS 将进入 SPSS 17.0 for Windows 安装向导界面 1，如图 1.2 所示。用户还可以根据自己的需要选择其他欲安装的模块。

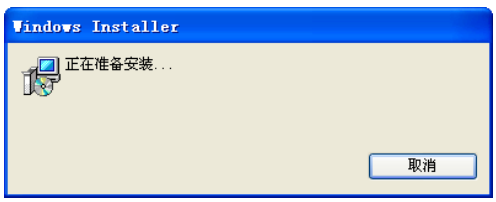


图 1.2　SPSS 17.0 for Windows 安装向导界面 1

03 用户可指定安装的许可密钥类型，如图 1.3 所示。

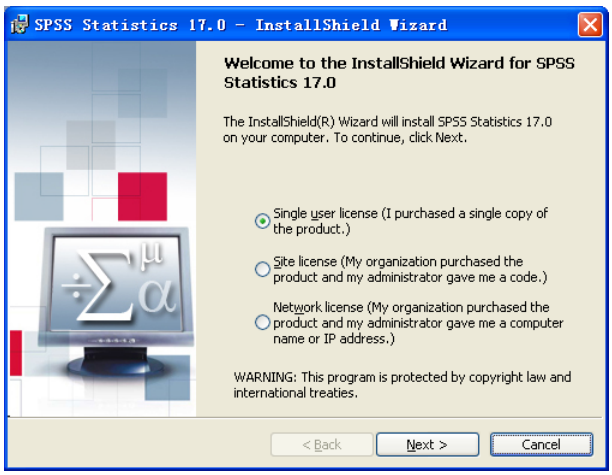


图 1.3　SPSS 17.0 for Windows 安装向导界面 2

- Single user license（个人用户许可）：针对个人客户，需要购买注册码。
- Site license（输入许可密钥）：针对团体客户，组织统一购买注册码，并由管理员给个人电脑分配注册码。
- Network license（网络许可）：针对团体客户，管理员分配电脑用户名和 IP 地址。

在选定安装的许可密钥类型后，单击“Next（下一步）”按钮。

04 签订许可协议。在详细阅读 SPSS 用户许可协议后，签订协议，如图 1.4 所示。

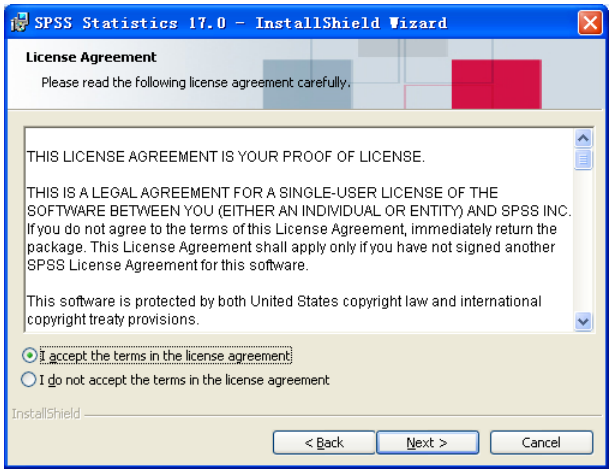


图 1.4　SPSS 17.0 for Windows 安装向导界面 3

- I accept the terms in the license agreement（我接受许可协议条款）：选中后，单击“Next（下一步）”按钮，即可进入下一界面。

- I do not accept the terms in the license agreement（我不接受许可协议条款）：选中后，“Next（下一步）”按钮将不能单击，无法进入下一界面，只能退出安装。

选中“（I accept the terms in the license agreement 我接受许可协议条款）”，单击“Next（下一步）”按钮，进入下一界面。

05 阅读 SPSS17.0 for Windows 说明，如图 1.5 所示。阅读完毕，单击“Next（下一步）”按钮，进入下一界面。

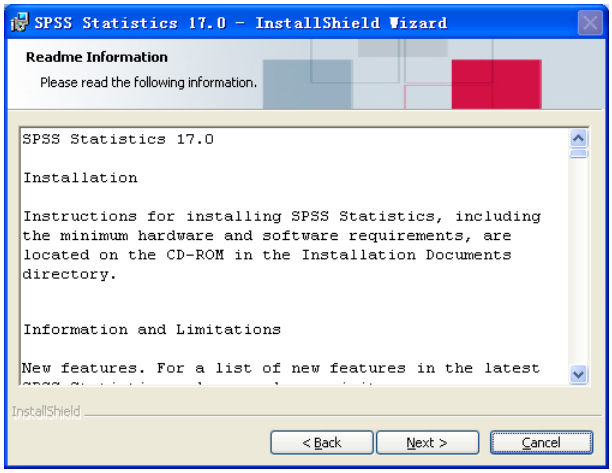


图 1.5　SPSS 17.0 for Windows 安装向导界面 4

06 在如图 1.6 所示界面输入用户姓名（User Name）和单位名称（Organization）后，单击“Next（下一步）”按钮，进入下一界面。

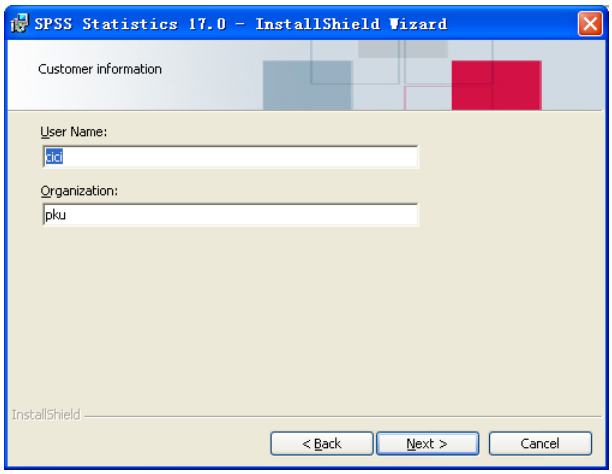


图 1.6　SPSS 17.0 for Windows 安装向导界面 5

07 选择 SPSS 17.0 for Windows 的安装位置，如图 1.7 所示。默认的安装目录为 C 盘

下 Program Files 新创建的 SPSS 文件夹。单击右端的“Change（更改）”按钮，可以改变安装位置。确定安装位置后，单击“Next（下一步）”按钮，进入下一界面。

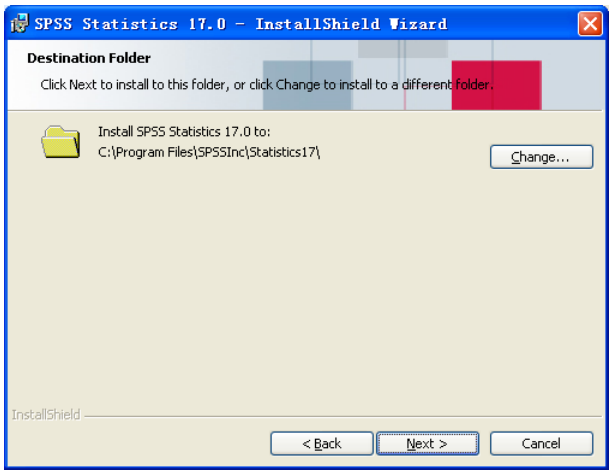


图 1.7　SPSS 17.0 for Windows 安装向导界面 6

08 单击“Install（安装）”按钮即可开始安装程序，如图 1.8 所示。

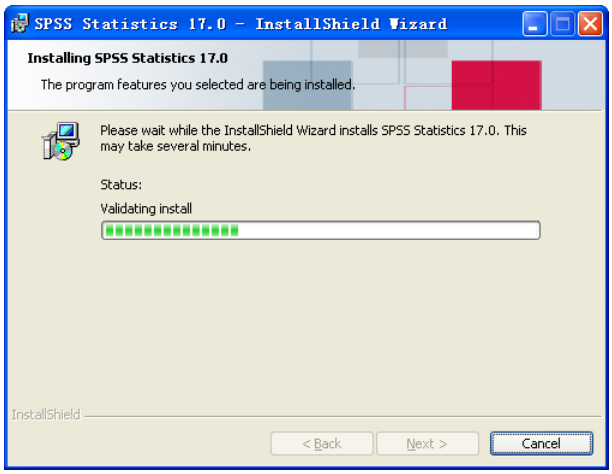


图 1.8　SPSS 17.0 for Windows 安装向导界面 7

09 许可认证，并注册 SPSS 17.0 for Windows，如图 1.9 所示。

通过 SPSS 许可认证的方法主要有以下两种：

- 第一种方法是根据授权的许可注册码，输入完成注册，此时 SPSS 17.0 可永久正常使用；
- 第二种方法是通过在 SPSS 主页（spss.com）注册，获得试用期注册码，这种情况下，用户将拥有 14 天的试用期。过期后，SPSS 将无法使用。

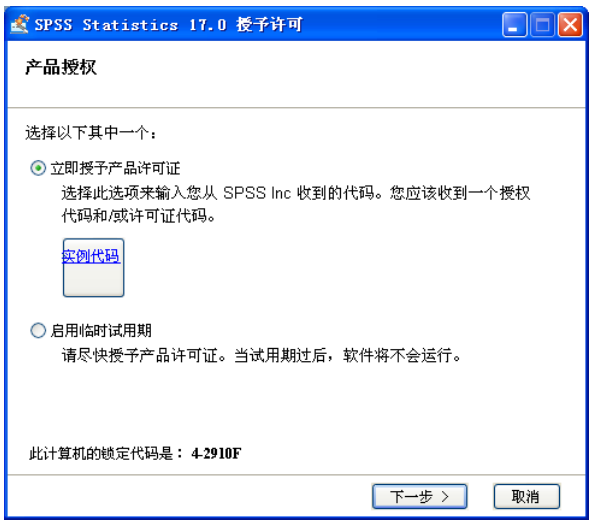


图 1.9　SPSS 17.0 for Windows 安装向导界面 8

完成许可认证后，单击“OK（完成）”按钮，即完成了 SPSS 17.0 for Windows 的安装。用户可以开始 SPSS 的使用之旅。

1.4　SPSS 的主要界面

1.4.1　SPSS 的启动

SPSS 安装完毕后，系统会自动在 Windows“开始”菜单下创建快捷方式。打开“开始”菜单，在“SPSS for Windows”下选中“SPSS 17.0 for Windows”命令并单击，即可启动 SPSS。

当用户运行 SPSS 软件后，计算机屏幕上出现一个“SPSS 启动操作”对话框，如图 1.10 所示。

在“SPSS 启动操作”对话框中，用户可以选择打开数据的方式。对话框中包括一个六选一的单选框和一个复选框，分别说明如下。

◆ **“您希望做什么？**

- 运行教程（Run the Tutorial）：单击后，SPSS 将打开帮助教程，在教程中，用户可选择不同模块的帮助说明进行针对性的辅导。
- 输入数据（Type in data）：需要手动输入数据，建立新的数据文件时可选择此项。选中后，即进入空白的 SPSS 数据编辑窗口。
- 运行现有查询（Run an existing query）：选中后，可以选择查询文件（.spq）的位置，并可单击打开。
- 使用数据库向导创建新查询（Create new query using Database Wizard）：选中后，将进入数据库向导对话框，根据向导指示可以建立新查询文件。

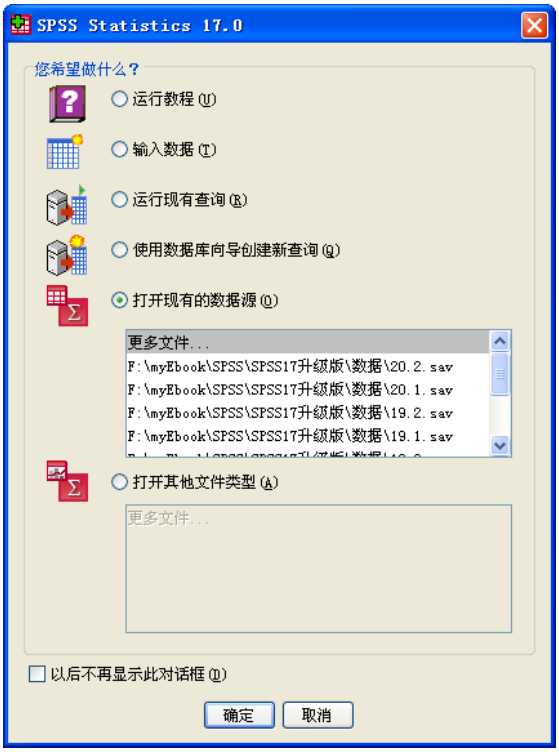


图 1.10　SPSS 启动操作对话框

- 打开现有的数据源（Open an existing date source）：选中后，选择现存数据源（.sav）的位置，并可单击打开文件。
- 打开其他文件类型（Open another type of file）：选中后，选择其他类型文件的位置，并可单击打开文件。SPSS 可以打开的数据类型包括：SPSS（.sav、.sys、.syd、.por）、Excel（.xls）、Lotus（.w）、Text（.txt）等。

◆ **复选项**

以后不再显示此对话框（Don't show this dialog in the future）：选中该复选项后，下次启动 SPSS 时将不再显示“SPSS 启动操作”对话框，直接显示数据编辑窗口。

选择任一选项，打开文件后，用户将进入 SPSS 的数据编辑窗口。

1.4.2　SPSS 的数据编辑窗口

SPSS 的主要窗口包括数据编辑窗口和结果输出窗口。

数据编辑窗口结构与 Microsoft Office 的 Excel 类似，主体的数据编辑区都是单元格的形式，但 SPSS 的数据编辑拥有更多的统计功能，其数据编辑窗口如图 1.11 所示。

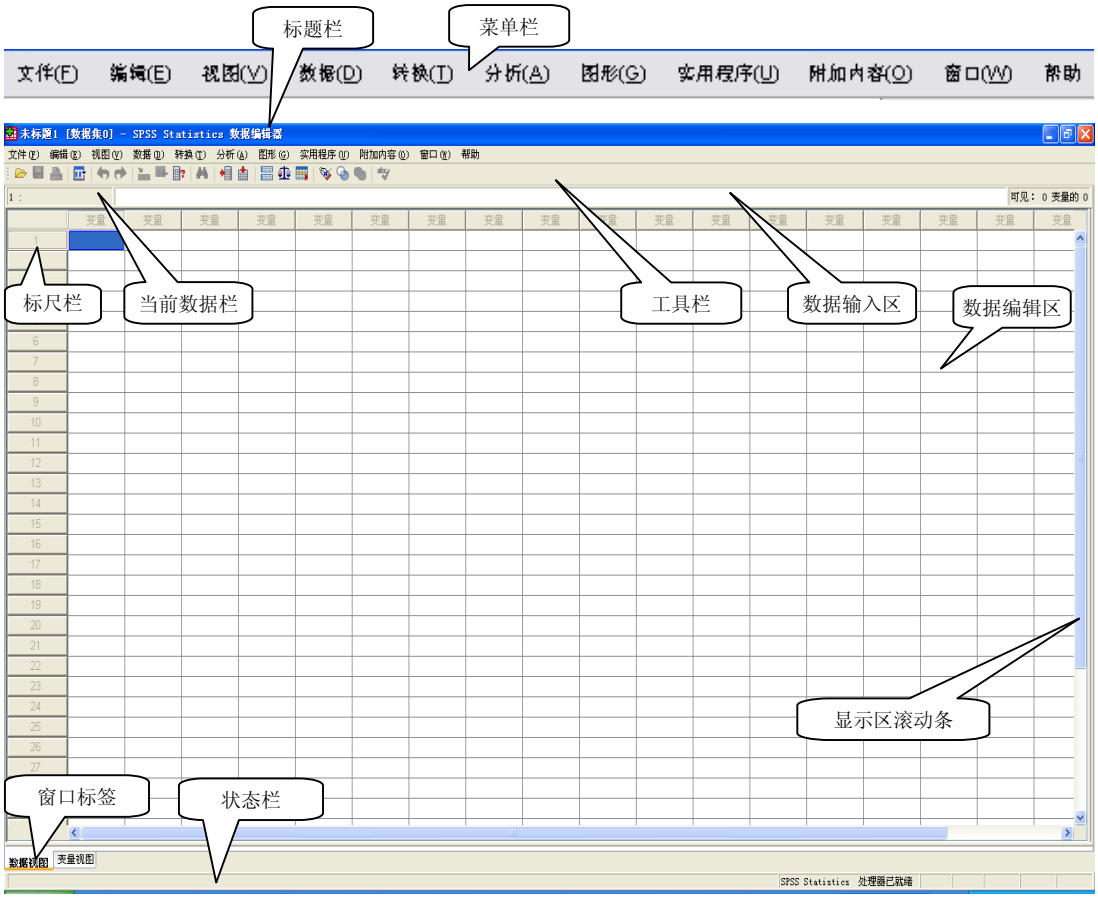


图 1.11　SPSS 数据编辑窗口数据视图

（1）标题栏：显示文件名，当文件尚未保存时，标题栏上显示为“未标题（Untitled）”。标题栏右端有三个按钮，分别可以“最小化”、“最大化”、“关闭”SPSS 数据编辑窗口。

（2）菜单栏：单击菜单选项即可激活菜单，此时弹出下拉式子菜单，用户可根据自己的需求单击子菜单的选项，完成特定的功能。菜单栏所包含的菜单选项如下。

“文件（File）”菜单：实现有关文件的调入、存储、显示和打印等。

“编辑（Edit）”菜单：实现有关文本内容的选择、复制、剪贴、寻找和替换等。

“视图（View）”菜单：用户可对数据编辑窗口的各栏目是否显示进行选择。

“数据（Data）”菜单：实现有关数据变量定义、数据格式选定、观察对象的选择、排序、加权、数据文件的转换、连接、汇总等。

“转换（Transform）”菜单：在数据转换处理菜单下，可以实现有关数值的计算、重新赋值、缺失值替代等。

“分析（Analyze）”菜单：应用一系列的统计方法输出分析数据、图表等。

“图形（Graphs）”菜单：制作相关统计图。

实用程序菜单：用户可以进行命令解释、字体选择、文件信息、定义输出标题、窗口设计等。

“附加内容（Add-ons）”菜单：用户可以应用 SPSS 的辅助软件，如 Amos 等进行深入分析。

“窗口（Windows）”菜单：用户可以进行窗口的管理，包括窗口的排列、选择、显示等。

“帮助（Help）”菜单：帮助用户调用、查询、显示文件等。

（3）工具栏：单击工具栏上的工具选项即可激活特定的功能，如图 1.12 所示。SPSS 工具栏所包含的工具选项都是在 SPSS 统计分析中最常用的，从左到右依次如下。



图 1.12　SPSS 数据编辑窗口工具栏

打开（Open File）：可打开数据文件、语句文件、其他类型的文件（如 Excel）等。

保存（Save File）：对数据进行编辑修改后，可以将修改后的内容进行保存。

打印（Print）：打印输出数据编辑区的报表。

恢复对话框（Dialog Recall）：单击后，可显示最近打开的对话框，重新对对话框进行编辑。如果打开文件后，从未打开任何对话框，则该功能不可用。

撤销（Undo）：可撤销上一步或上几步的错误操作。

恢复（Redo）：撤销操作后，若想还原，可单击该按钮。

到达记录（Goto case）：单击后，将打开“到达记录（Goto case）”对话框，在对话框内输入数字，可立即到达当前单元格所在变量（列）的记录（行）处。

到达变量（Goto Variables）：单击后，转换至变量编辑窗口。

变量（Variables）：单击后，将打开“变量（Variables）”对话框，对话框内将显示全部数据变量的名称、标签、类型等形式。

查找（Find）：单击后，将打开“查找（Find）”对话框，在对话框内输入要查找的内容，即可找到该内容所在的单元格。

插入个案（Insert cases）：单击后，将在光标所在单元格的上方插入一行记录，可在该行中输入新的个案。

插入变量（Insert variable）：单击后，将在光标所在单元格的左方插入一列变量，可在该列中输入新的变量值。

拆分文件（Split File）：单击后，将打开“拆分文件（Split File）”对话框，可在该对话框中对文件进行分组，进而拆分文件。

加权个案（Weight cases）：单击后，将打开“加权个案（Weight cases）”对话框，可在该对话框中，选择频数变量对个案进行加权求和。

选择个案（Select cases）：单击后，将打开“选择个案（Select cases）”对话框，

可在该对话框中，对所有个案设定条件、范围及样本的随机性，进而筛选出所有满足条件的个案。

数值标签（Value Labels）：如果已对变量进行标签设定（下文将进行说明），单击该按钮，变量将被所对应的数值标签替代，再次单击该按钮，即可恢复显示。

使用集合（Use variable sets）：将变量分组定义为集合后，可单击该按钮，在打开的“使用集合（Use variable sets）”对话框中，选择在数据编辑区中显示的变量集合。

显示所有变量（Show all variables）：单击后，将在数据编辑区内显示所有变量。

（4）当前数据栏：以“记录名：变量名”的形式，说明当前单元格所在的位置。

（5）数据输入区：显示当前单元格的数据内容，在数据输入区可以直接修改数据。

（6）标尺栏：在标尺栏可以调节行、列的宽度。标尺栏有纵向标尺栏和横向标尺栏两种，横向标尺栏显示数据变量，纵向标尺栏显示数据顺序（如时间顺序）。

（7）数据编辑区：数据编辑区是 SPSS 的主要操作窗口，是一个二维平面表格，用于对数据进行各种编辑。SPSS 的数据编辑区与 Excel 的数据编辑区类似，都是以行、列的形式组成单元格，并可编辑单元格，进而修改整个报表数据。数据编辑区的行称为“记录”或“个案”，列称为“变量”。

（8）显示区滚动条：通过调节滚动条，可以显示报表不同位置的数据。

（9）窗口标签：进入 SPSS，系统显示数据编辑窗口的数据视图（Data View）。数据编辑窗口可以显示两张表，分别是数据视图（Data View）和变量视图（Variable View）（如图 1.13 所示）。在数据视图（Data View）中，用户可以直接输入观测数据值或存放数据，表的左端列边框显示个案的序号，最上端行边框显示变量名。变量视图（Variable View）用来定义和修改变量的名称、类型及其他属性。通过单击数据编辑窗口下端的两个同名窗口标签按钮可以实现两者的相互切换。

在变量视图（Variable View）中，与数据视图（Data View）不同，每一行描述一个变量，每一列描述该变量的属性。变量属性从左到右依次是：

- 变量名（Name）：变量名必须以字母、汉字或@开头，总长度不超过 8 个字符，即最多可容纳 4 个汉字或 8 个英文字母。英文字母不区分大小写，但最后一个字符不能是句点。
- 变量类型（Type）：变量类型有 8 种，最常用的是数值型（Numeric）变量。其他常用的类型包括：字符型（String）、日期型（Date）、逗号型（Comma，即数值隔 3 位数加一个逗号）等。
- 变量宽度（Width）：变量中取值所占的宽度，默认的变量宽度为 8。
- 小数点后位数（Decimals）：限制数值型、逗号型等变量类型数值的小数点后的位数，默认的小数点后位数为 2。字符型不能限制小数点后的位数。
- 变量标签（Label）：可通过变量标签设定变量含义的详细说明。
- 变量值标签（Values）：可通过变量标签设定变量取值具体含义的详细说明。

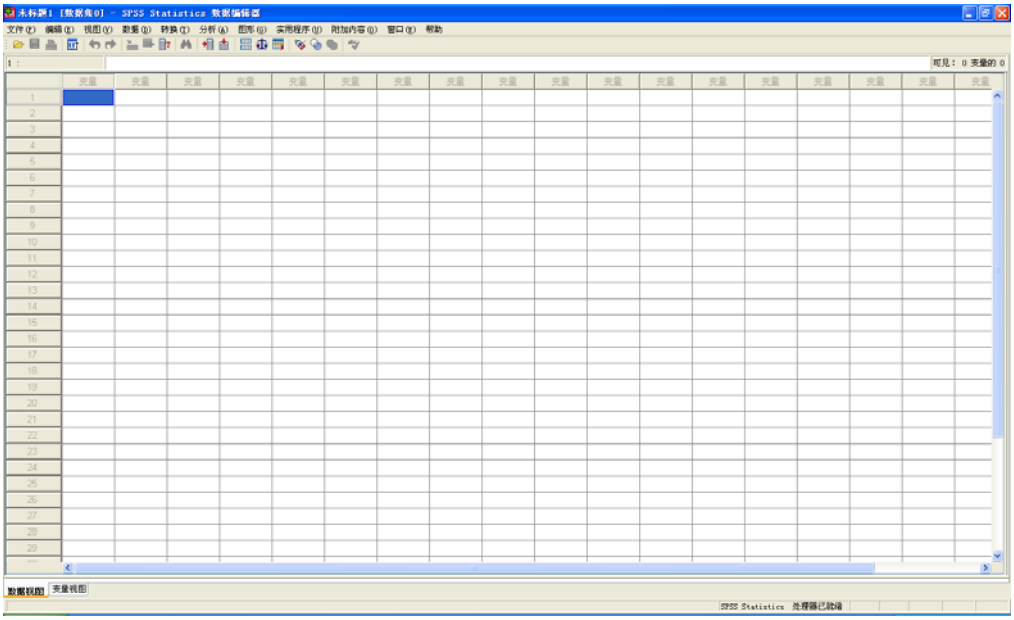


图 1.13　SPSS 数据编辑窗口变量视图

- 缺失值（Missing）：选择缺失值的处理方式，包括不处理、离散缺失值等。
- 列宽（Columns）：变量在数据视图（Date View）中所显示的列宽，默认的列宽为 8。
- 对齐格式（Align）：数据默认的对齐格式为右对齐。
- 测度方式（Measure）：系统给出的数据的测度方式包括名义尺度、定序尺度和等间距尺度三种，默认的测度方式为等间距尺度。

如果输入变量名后回车，系统将给出变量的默认属性。如果不定义变量的属性，直接输入数据，系统将设定变量名为默认的 Var00001、Var00002 等。

定义了变量的各种属性后，单击 Date View，回到数据视图中，就可以直接在表中录入数据。输入数据后可以单击“保存（Save）”按钮或者“另存为（Save as）”按钮作为数据文件保存。

（10）状态栏：显示 SPSS 的状态，如果显示的是“SPSS Professor is ready”，则 SPSS 可正常使用，并且为等待用户操作的状态。

1.4.3　SPSS 的结果输出窗口

结果输出窗口主要显示统计分析的结果、图表以及信息说明等，此窗口的内容以结果文件 Output（.spo）的形式保存，如图 1.14 所示。

SPSS 结果输出窗口的布局与数据编辑窗口类似，而结果浏览区域和 Windows 资源管理器的结构也几乎相同。下面将对存在差异的工具栏、大纲视图和结果输出区进行说明。

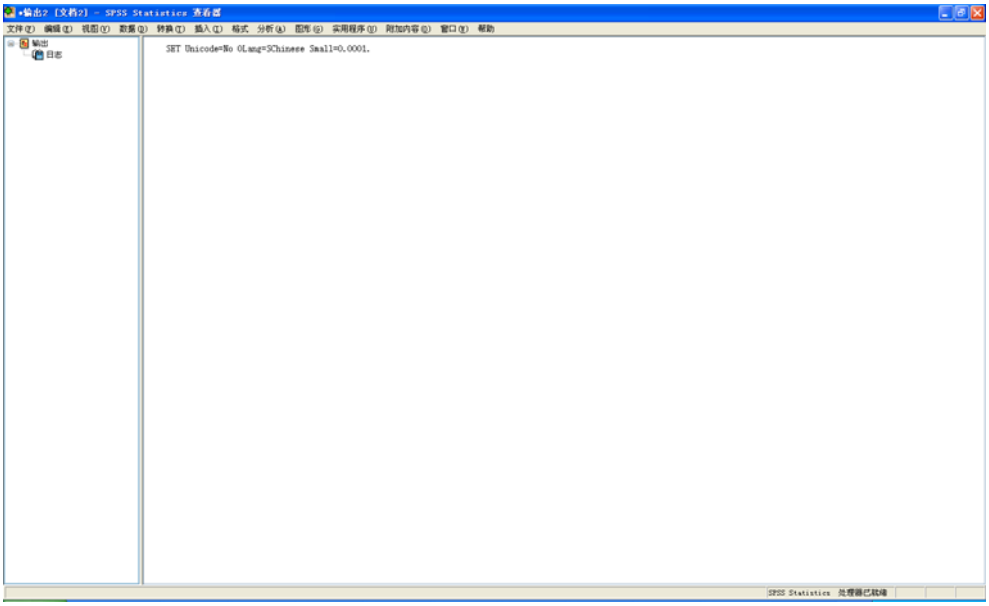


图 1.14　SPSS 结果输出窗口

（1）工具栏

单击工具栏上的工具选项即可激活特定的功能， SPSS 结果输出窗口的工具栏如图 1.15 所示。



图 1.15　SPSS 结果输出窗口工具栏

工具栏从左到右依次为：

打开（Open File）：可打开结果文件 Output（.spo）。

保存（Save File）：对输出结果进行保存。

打印（Print）：打印结果输出区的报表、图形。

打印预览（Print Preview）：打印前对结果输出区的报表、图形进行预览、页面编辑。

导出（Export）：导出结果输出区的报表、图形，结果可以导出到 html 文件（.htm）、文本文件（.txt）、Excel 文件（.xls）、Word 文件（.doc）、PowerPoint 文件（.ppt）中。

恢复对话框（Dialog Recall）：单击后，可显示最近打开的对话框，重新对对话框进行编辑。如果打开文件后，从未打开任何对话框，则该功能不可用。

撤销（Undo）：可撤销上一步或上几步错误操作。

恢复（Redo）：重新执行用户操作：

到达数据（Goto Data）：单击后，转换至数据编辑窗口。

到达记录（Goto case）：单击后，将打开“到达记录（Goto case）”对话框，在对

话框内输入数字，即可立即到达数据编辑窗口当前单元格所在变量（列）的记录（行）处。

到达变量（Goto Variables）：单击后，转换至变量编辑窗口。

变量（Variables）：单击后，将打开“变量（Variables）”对话框，对话框内将显示数据编辑窗口全部数据变量的名称、标签、类型等形式。

使用集合（Use variable sets）：将变量分组定义为集合后，可单击该按钮，在打开的“使用集合（Use variable sets）”对话框中，选择在数据编辑窗口显示的变量集合。

显示所有变量（Show all variables）：单击后，将在数据编辑窗口内显示所有变量。

选定最后输出（Select Last Output）：单击后，自动选中结果输出区最后的一个输出结果。

关联自动脚本（Associate Autoscript）：单击后，将输出结果和自动脚本关联在一起。

创建/编辑自动脚本（Create/Edit Autoscript）：单击后，将调出脚本页面，用户可以创建或编辑自动脚本。

运行脚本（Run script）：单击后，将运行现有脚本，输出结果。

候选窗口（Designate Window）：单击后，当前窗口被设定为候选结果输出窗口，之后的输出结果将出现在当前窗口中。如果当前窗口已是候选窗口，此按钮将不可用。如果只打开一个结果输出窗口，该窗口将被自动设定为候选窗口，此按钮也不可用。

升级（Promote）：选中结果输出区的输出结果后，单击该按钮，可以向前调整该结果的层级。层级关系将在大纲视图的树形图中展现。

降级（Demote）：选中结果输出区的输出结果后，单击该按钮，可以向后调整该结果的层级。层级关系将在大纲视图的树形图中展现。

展开（Expand selected Outline items）：在大纲视图中选定非最低层级的结果输出后，单击该按钮，将展开该层级下所有低层级的图标。

折叠（Collapse selected Outline items）：在大纲视图中选定非最低层级的结果输出后，单击该按钮，将不显示该层级下所有低层级的图标。

显示（Show selected items）：在大纲视图中选定非最低层级的结果输出后，单击该按钮，将显示该层级下被隐藏的输出结果。

隐藏（Hide selected items）：在结果输出区选定输出结果后，单击该按钮，可将该结果隐藏。

插入新题目（Insert Heading）：在结果输出区选定输出结果后，单击该按钮，可在该结果下方插入新题目，新题目将在大纲视图中显示，并可对新题目进行编辑修改。

插入新标题（New Title）：在结果输出区选定输出结果后，单击该按钮，可在该结果下方插入新标题，新标题可直接在结果输出区进行编辑修改。

插入文本（New Text）：在结果输出区选定输出结果后，单击该按钮，可在该结果下方插入新文本，新文本可直接在结果输出区进行编辑修改，用于补充说明上方输出图表。

（2）结果浏览区域

工具栏以下的窗口被纵向一分为二，左侧是大纲视图（Outline View，又称结构视图），

右侧则显示详细的统计结果（统计表、统计图和文本结果）。两侧的元素是完全一一对应的，即选中一侧的某元素，在另一侧该元素也会被选中。

大纲视图的各个元素如下。

 大纲图标，代表一段或整个输出结果，含下级元素，单击左侧的减号就可以将下级元素折叠，折叠后减号变为加号，图标则变为 。

 运行记录图标，代表系统操作产生的一段运行记录。

 警告图标，代表输出结果中的系统警告。

 注解图标，代表系统自动产生的注解，默认情况下注解内容在输出结果中是隐藏的。

 标题图标，代表输出标题。

 页标题图标，代表输出标题，较少出现。

 表格图标，代表输出结果中的统计表（Pivot table，数据透视表）。

 统计图图标，代表统计图。

 文本图标，代表文本输出结果。

 交互式统计图图标，代表交互式统计图。

 统计地图图标，代表统计地图。

在大纲视图中，单击图标会选中所代表的一块或一段输出结果，双击图标可以让对应的输出结果在显示、隐藏间切换，选中图标后单击图标的名称便可更改图标名称。

在结果输出区中，双击输出结果，可以对输出的结果进行格式、样式的编辑。

1.5　本章小结

本章主要介绍了 SPSS 17.0 for Windows 中文版的特点、运行方式、安装步骤及用户界面等内容。通过对本章的学习，读者可以对 SPSS 的数据编辑窗口和结果输出窗口有大体的认识，为以后逐步熟悉和掌握 SPSS 的操作打下基础。

第2章 数据的基本操作

在应用 SPSS 进行数理统计分析前，首先要建立数据文件，并对所建立的数据文件进行简单的编辑和整理，在此基础上，之后的统计分析才是有效的。SPSS 能够与多种数据库或数据文件进行连接，在很大程度上为建立数据文件提供了方便。同时，SPSS 所提供的强大的数据编辑功能能够统一地对数据进行分组、转换等操作，简化了复杂的数据处理流程。

SPSS 和大多数应用软件一样，数据文件管理的主要功能都在“文件（File）”菜单下，即在“文件（File）”菜单下可以实现新建文件、保存文件等功能，而对于数据编辑，SPSS 则提供了“编辑（Edit）”菜单、“数据（Data）”菜单和“转换（Transform）”菜单。

本章学习要点：

- 建立数据文件；
- 定义数据类型；
- 数据的排序；
- 数据文件的合并；
- 数据的转换；
- 数据的分组；
- 数据的分类汇总；
- 保存数据文件；
- 导出数据文件。

2.1 建立数据文件

在进行数据处理之前，首先要建立数据文件。建立数据文件的方式主要有四种：人工

输入数据、打开 Excel 等数据文件、使用数据库查询和导入数据文件。下面将分别通过这四种方式来建立数据文件。

2.1.1　输入数据建立数据文件

输入数据建立数据文件适用于数据刚收集完成，还未输入任何数据的文件；此时，需要在 SPSS 中手动输入数据，建立数据文件。SPSS 所提供的手动输入数据建立数据文件的方式极为简便。启动 SPSS 后，系统自动生成一个空数据文件，此时，只要定义变量类型，输入数据，然后保存即可。

如果需要在已有数据的 SPSS 数据编辑窗口中新建数据文件，则需要将原有的数据进行保存，然后单击“文件（File）”菜单下的“新建（New）”→“数据（Data）”命令，即可新建一个数据文件。如图 2.1 所示。



图 2.1　“文件（File）”菜单下的“新建（New）”选项

同样，如果在“文件（File）”菜单下的“新建（New）”选项中，选择“语法（Syntax）”、“输出（Output）”、“脚本（Script）”，则可分别打开对应窗口，进而分别建立新的.sps、.spo、.rtf、.sbs 文件。

2.1.2　直接打开其他格式的数据文件

如果数据搜集完毕，已经输入并储存为其他类型的数据文件，那么在用 SPSS 进行数

据处理前，就不需要重新进行数据输入，直接打开其他格式的数据文件即可。SPSS 能够与多种数据文件进行连接，简化烦琐的输入步骤。

单击“文件（File）”菜单下的“打开（New）”→“数据（Data）”，即可打开一个其他格式的数据文件。如图 2.2 所示。

单击“数据（Data）”后，将弹出“打开数据（Open Data）”对话框，如图 2.3 所示。在文件类型中，用户单击右端的小箭头，单击数据文件对应的数据类型，并通过选择数据文件所在位置及选定数据文件，然后单击“打开（Open）”按钮，即可打开其他格式的数据文件。SPSS 可以打开的数据类型包括：SPSS（.sav、.sys、.syd、.por）、Excel（.xls）、Lotus（.w）、Text（.txt）等。



图 2.2 “文件（File）”菜单下的“打开（Open）”选项

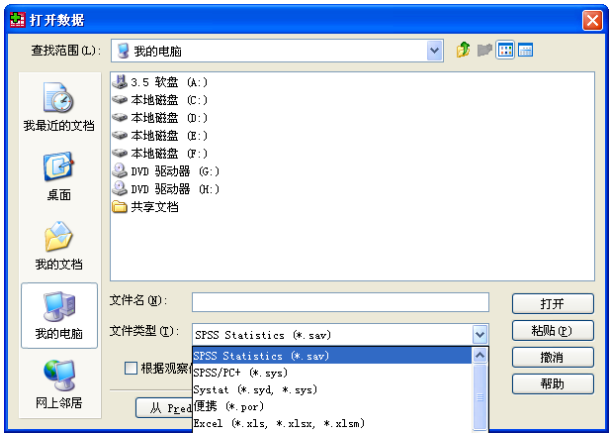


图 2.3 “打开数据（Open Data）”对话框

2.1.3 使用数据库查询建立数据文件

除了可以直接打开许多类型的数据文件外，SPSS 还提供了另一个适用范围更广、但使用也更为专业的数据接口——数据库查询。SPSS 使用了 ODBC（Open Database Connectivity）数据接口，该接口能被大多数数据库软件和办公软件（如 Microsoft Office）支持，通过 ODBC 应用程序可以直接访问以结构化查询语言（SQL）作为数据访问标准的数据库管理系统。由于 SPSS 17.0 for Windows 中文版可以直接打开除 Excel 2007 外的所有 Excel 数据文件，因此，数据库查询接口对于打开 Excel 文件的用处不是很大。

单击“文件（File）”菜单，选择“打开数据库（Open Database）”下的“新建查询（New Query）”命令，如图 2.4 所示，系统会弹出“数据库向导（Database Wizard）”窗口，如图 2.5 所示，窗口中将列出所用机器上已安装的所有数据库驱动程序，选中所需的数据源，

根据向导的提示即可将数据导入 SPSS。

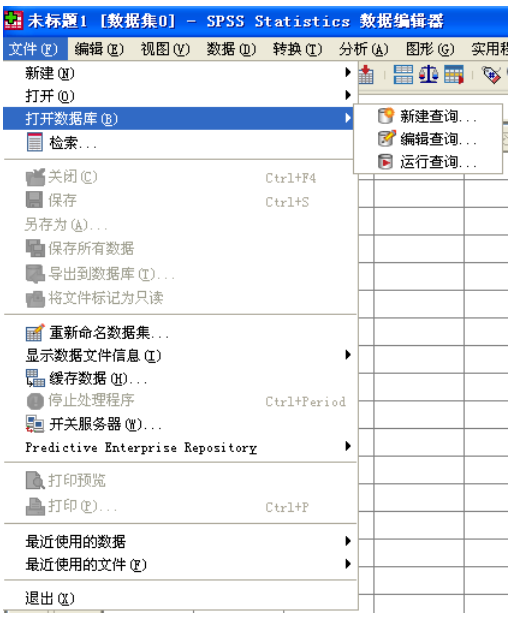


图 2.4 “文件（File）”菜单下的“打开数据库（Open Datebase）”选项

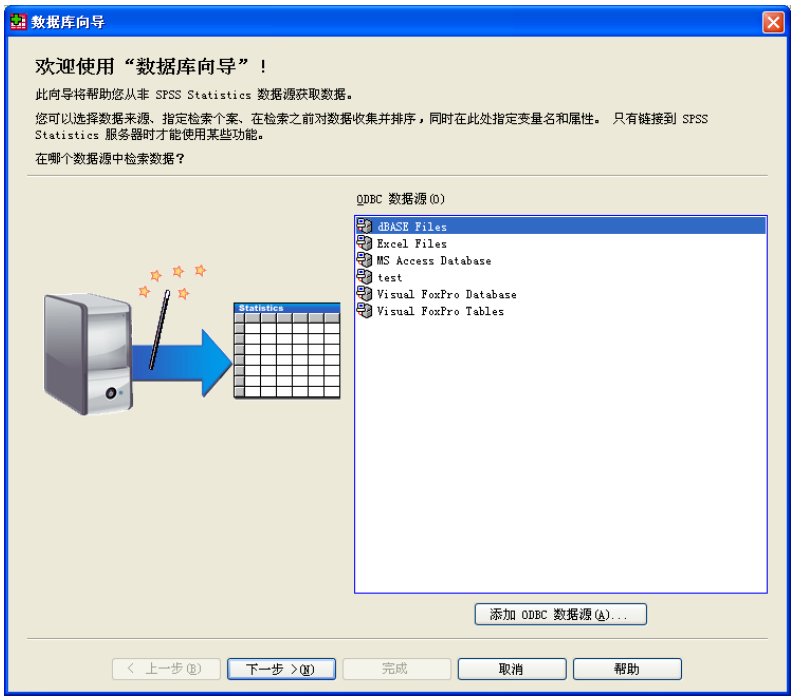


图 2.5 “数据库向导”对话框

2.1.4 导入文本文件建立数据文件

SPSS 提供了从文本文件(.txt)导入数据建立数据文件的渠道。实际上，该功能在 SPSS 中已被整合到“打开文件（Open File)”对话框中。之所以在菜单上保留该条目有两个原因：一是读入纯文本的情况非常普遍，保存在菜单上更为醒目；二是为了和 SPSS 老版本的使用保持兼容。

打开“文件（File)”菜单，选择“检索（Read Text Data)”命令，如图 2.6 所示，系统就会弹出“打开数据（Open File）对话框，操作和 2.1.2 节所述完全相同。

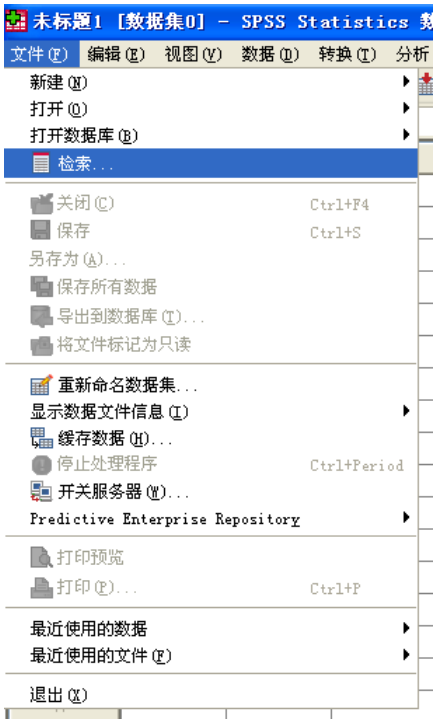


图 2.6 “文件（File)”菜单下的“检索（Read Text Data)”选项

2.2 编辑数据文件

在数据视图（Data View）中，用鼠标单击数据视图左边框的个案序号，对应行就会被选中，用鼠标单击上边框的变量名，对应列就会被选中。与其他 Windows 中的操作类似，用鼠标选中行、列、部分单元格后，单击鼠标右键，可以对它们进行复制（Copy)、删除（Delete)、剪切（Cut)、清除（Clear）等操作。

数据输入完毕后，需要对已经存在的数据进行修改和整理，这就需要用到 SPSS 的数据编辑功能。下面介绍几种常用的数据编辑功能。

2.2.1　输入数据

SPSS 和其他常用统计软件一样，也提供了直接输入数据的功能。SPSS 数据文件中的数据是在 SPSS 数据编辑窗口中以表格的形式进行输入的。数据编辑窗口中的黑色边缘的单元格即为当前数据单元格，表示用户正在对单元格进行数据输入或数据修改。因此，在数据输入时，用户首先应确定哪个单元格为当前单元格。确定当前单元格后，只要将鼠标指针定位在需要进行数据编辑的单元格上，然后单击鼠标左键即可。

数据输入时，可以逐行输入或者逐列输入。逐行输入时，只需输完一个数据后，按回车键或者↓键，当前单元格将跳到本行的下一列变量上。逐列输入数据时，只需输完一个数据后，按 Tab 键或者→键，当前单元格将跳到同一列变量的下一行个案上。

SPSS 数据界面最大的优势就是支持鼠标的拖放操作，以及复制、粘贴等命令，这样就大大提高了用户的数据输入速度。这里，需要提及几个比较常用的数据输入的技巧。

1．连续多个相同值的输入

在直接输入数据时，如需要输入连续多个“1”，就可以用以下方法：

（1）在第一格内输入“1”并按回车键；

（2）回到第一格并单击右键，选择“复制（Copy）”；

（3）用拖放方式选中所有应输入“1”的单元格，单击右键并选择“粘贴（Paste）”，则所有选中的单元格就都会被刚才复制的“1”填充。

2．将 Excel 数据直接复制到 SPSS 中

Excel 打开原数据后，SPSS 将不能打开正在使用的 Excel 文件。这时，如果 Excel 工作簿中的数据量较少，可以直接用复制、粘贴的方法将数据导入 SPSS 中，而无需关闭 Excel 原数据文件，再使用 SPSS“文件（File）”菜单下的“打开数据（Open File）”选项。具体操作步骤如下：

01 在 Excel 工作簿中选中所有需要复制的数据（不包括变量名）；

02 单击右键，选择“复制（Copy）”（或直接按键盘上的 Ctrl+C 组合键）；

03 切换到 SPSS 中，将需要复制的单元格选定为当前单元格，然后执行“粘贴（Paste）”命令，数据就会全部转入 SPSS 中，此后再定义相应的变量即可。

需要注意的是，这种直接复制的方式只适用于数据量较少的情况，如果数据量较大，SPSS 将只复制 Excel 在一个页面（即窗口上显示的当前页面）上的数据，此种情况下，关闭 Excel 文件，运行 SPSS“文件（File）”菜单下的“打开数据（Open）”选项，则更为简便。

2.2.2　定义数据属性

数据的属性主要包括每列数据变量的变量名、变量类型、变量宽度、变量名标签、变

量值标签、变量列格式、变量缺失值、变量量度方法等信息。使用 SPSS 时，大多数情况下需要直接定义新变量，此时，只需单击数据编辑窗口页面左下方的“变量视图（Variable View）”标签就可以切换到变量定义界面开始定义新变量，如图 2.7 所示。在 1.4.2 节中，已经简单介绍过变量视图，这里将对变量视图的每一列选项进行详细地介绍。如果输入变量名后按回车键，则系统将给出变量的默认属性。如果不定义变量的属性，直接输入数据，系统将设定变量名为默认的变量名。

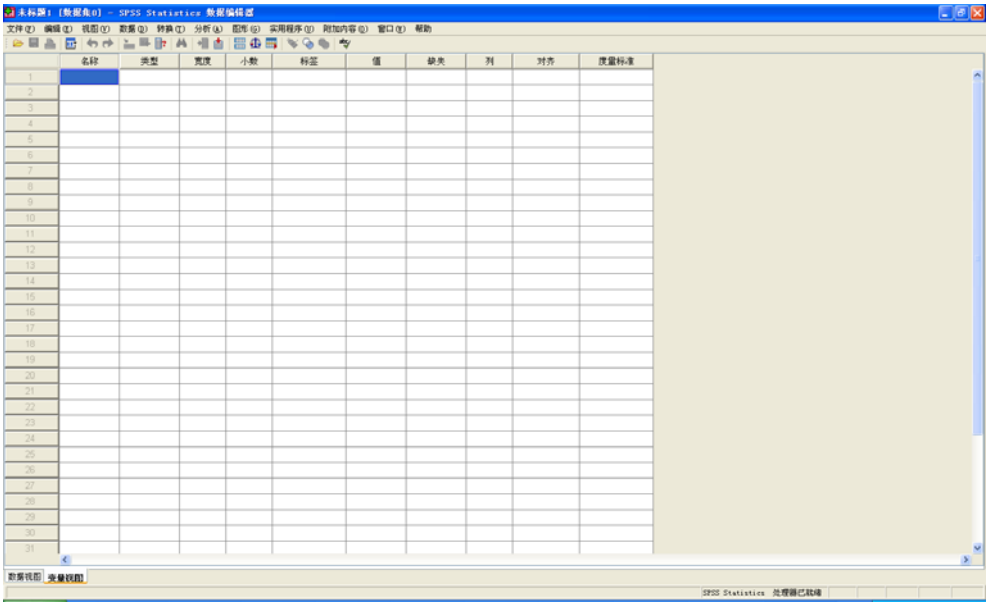


图 2.7　SPSS 数据编辑窗口变量视图

变量名（Name）：变量名是变量存取的唯一标志。单击“名称”列项的单元格，即可直接输入变量名。

SPSS 对于变量名的命名规则有以下要求：

（1）总长度不超过 8 个字符，即最多可容纳 4 个汉字或 8 个英文字母；

（2）必须以字母、汉字或@开头，英文字母不区别大小写，但最后一个字符不能是句点；

（3）变量名不能与 SPSS 内部特有的具有特定含义的保留字符相同，如 ALL、BY、AND、NOT、OR 等；

（4）SPSS 默认的变量名是以 VAR 开头，后面补足 5 位数字，如 VAR00001。

如果变量名不符合变量名的命名规则，变量会自动给出错误提示信息。

变量类型（Variable Type）：SPSS 中，数据变量的类型有八种，其中有三种为基本的变量类型，数值型（Numeric）、字符型（String）和日期型（Date）。

单击“变量类型（Variable Type）”该列的单元格，单元格右端将出现 … 按钮，单击

该按钮，将弹出“变量类型（Variable Type）”对话框，如图 2.8 所示。选择变量类型后，对应的宽度（Width）单元格将显示默认的宽度和样式。

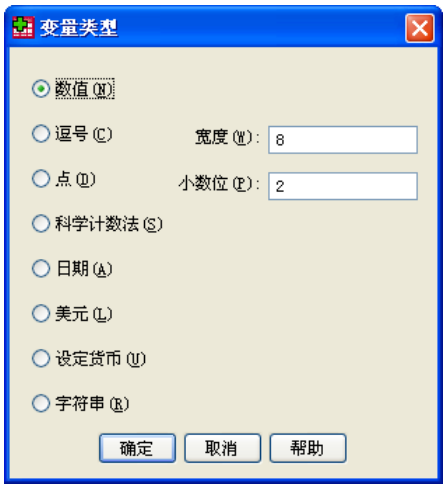


图 2.8 “变量类型（Variable Type）”对话框

（1）数值型（Numeric）

数值型是 SPSS 最常用的变量类型，数值型的数据是由 0~9 的阿拉伯数字和其他数字符号（如美元符号、逗号、圆点等）组成的。数值型有以下六种不同的表示方法。

- 标准数值型（Numeric）

标准数值型（Numeric）是系统默认的数据类型。标准数值型默认的最大数值显示宽度是 8 位，小数点后位数是 2 位。如果数据是整数，则它们最大的显示位数是 8 位正整数或 7 位负整数，如 12 345 678、–1 234 567。如果数据带有小数位数，则它们的总显示宽度将包括默认的 2 位小数和 1 位小数点，如 12 345.67、–1 234.56。

如果数值总宽度超过了 8 位，SPSS 将在显示总宽度为 8 的前提下，自动调整数据的显示，即对数据的最后几位进行四舍五入或者以科学计数法显示（当整数部分大于 8 位时），如 123 456.78 将显示为 123 456.8、1 234 567.89 显示为 1 234 568、123 456 789 显示为 1.23E+08、–123 456 789 显示为–1.2E+08。

需要说明的是，虽然 SPSS 对于数据显示宽度有一定的限制，但数据显示宽度并不影响实际数据的存储，也不影响数据的计量。用户可以根据具体情况对默认显示宽度进行修改。

- 逗号数值型（Comma）

逗号数值型（Comma）数据的整数部分从个位开始向前每 3 位以一个逗号分隔。逗号数值型的默认最大数值显示宽度是 8 位，小数点后位数是 2 位。逗号所占的位数将包括在总显示宽度之内，如 1,234.56。用户在输入逗号数值型数据时，可以不输入逗号，SPSS 将自动在相应位置上添加逗号。

同样地，如果数值总宽度超过了 8 位，SPSS 也将自动调整数据的显示。SPSS 首先会

去掉逗号的显示，然后按标准数值型的显示调整数据的显示，如 123,456.7 显示为 123456.7、1,234,567.8 显示为 1 234 568 等。

- 圆点数值型（Dot）

圆点数值型（Dot）数据的整数部分从个位开始向前每 3 位以一个圆点分隔，以逗号作为整数和小数部分的分隔符。逗号数值型的默认最大数值显示宽度是 8 位，小数点后位数是 2 位，如 1.234,56。用户在输入圆点数值型数据时，可以不输入圆点，SPSS 将自动在相应位置上添加圆点。

同样地，如果数值总宽度超过了 8 位，SPSS 将按照上述逗号数值型的显示调整规则自动调整数据的显示。

- 科学计数法型（Scientific Notation）

科学计数法型（Scientific Notation）是按照科学计数法来表示数据的方法之一。例如，1200 用科学计数法表示为 1.2E+03，其中的 E 表示以 10 为底，+03 表示正的 3 次方。科学计数法的默认最大数值显示宽度是 8 位，小数点后位数是 2 位，其中包括了整数前的正负符号、字母 E、其后的正负符号和具体数字。科学计数法一般用来表示较大或较小的数据。用户在使用科学计数法输入数据时，可以按标准数值型方式输入数据，SPSS 会自动进行转换。

- 美元数值型（Dollar）

美元数值型（Dollar）主要用来表示货币数据，具体数值前将附加美元符号 $。美元数值型（Dollar）的默认最大数值显示宽度是 8 位，小数点后位数是 2 位。美元数值型有很多显示格式，如 $#、$##、$###等，SPSS 以菜单的方式显示供用户选择。

- 用户自定义型（Custom Currency）

如果上述五种数值型数据表示方式仍不满足用户的需要，用户可自行定义最多四种表示方式，并且选择 SPSS 系统提供的五个名称 CCA、CCB、CCC、CCD、CCE 之一进行命名。实际上，SPSS 预先将 CCA、CCB、CCC、CCD、CCE 定义成逗号数值型，用户可以对它们进行重定义。

单击“编辑（Edit）”菜单下的“选项（Options）”命令，在弹出的“选项（Options）”对话框中，选择“货币（Currency）”界面，如图 2.9 所示。在“货币（Currency）”界面中，可以对数据表示方式进行重定义。

设定输入格式（Custom Output Formats）：用户选择 SPSS 系统提供的五个名称 CCA、CCB、CCC、CCD、CCE 之一作为重定义的输出格式。

样本输出（Sample Output）：显示输出格式的示例，包括正数和负数不同的显示。

所有值（All Values）：在“前缀（Prefix）”后面输入数值型数据的开始字符，在“后缀（Suffix）”后面输入数据的结尾字符。SPSS 默认的所有数值的前缀和后缀都是空格。

负值（Negative Values）：同样地，在“前缀（Prefix）”后面输入负数数值型数据的负号后的开始字符，在“后缀（Suffix）”后面输入负数数据的结尾字符。SPSS 默认的负数数值的前缀和后缀都是空格。

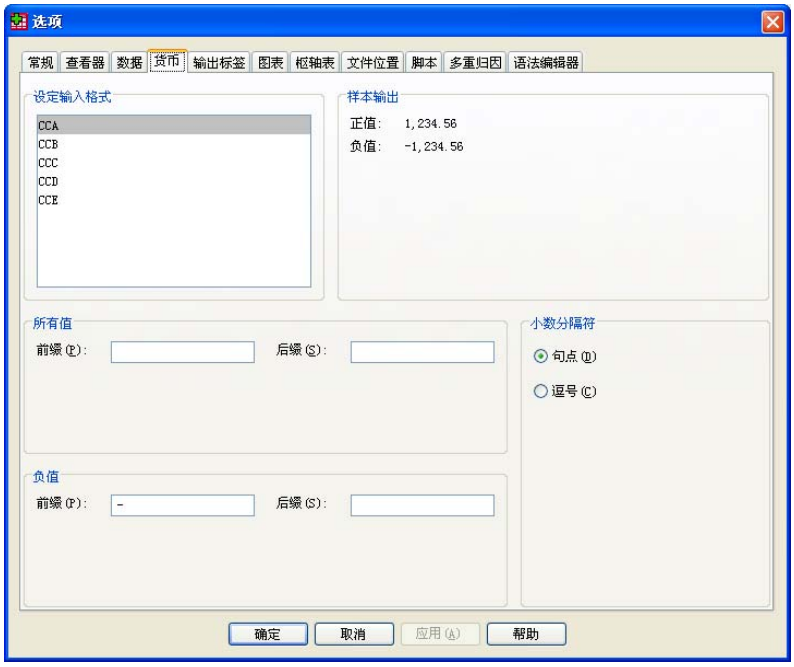


图 2.9 “选项（Options）”对话框下的“货币（Currency）”界面

小数分隔符（Decimal Separator）：指定是以句点（Period）还是以逗号（Comma）作为小数点。

对数值型数据表示方式进行重新定义后，单击“确定（OK）”按钮，即可存储自定义结果，用户可以在“变量类型（Variable Type）”界面选择新定义的数值型数据表示方式。

（2）字符型（String）

字符型（String）数据类型也是 SPSS 中较常用的数据类型，它是由字符串组成的。字符型数据的默认最大显示宽度是 8 个字符位，它不能够进行算术运算，并区分英文大小写字母。SPSS 数据处理过程中，通常将字符型数据用一对双引号括起来进行处理，因而在输入数据时无须输入双引号，否则 SPSS 也会将双引号作为字符型数据的一部分。

（3）日期型（Date）

日期型（Date）数据主要用来表示日期或者时间，例如生日等变量即可定义为日期型数据。日期型数据有很多显示格式，SPSS 以菜单的方式显示供用户选择。例如，dd-mmm-yyyy，dd 表示 2 个字符位的日期，-为数据分隔符，mmm 表示英文月份的缩写，yyyy 表示 4 个字符位的年份。

变量宽度（Width）：选择变量类型时，可以在“变量类型（Variable Type）”对话框中调整对应的宽度，默认的变量宽度为 8。用户也可以单击“变量宽度（Width）”该列的单元格，单元格将显示成 8，然后即可在单元格重新输入数值或者单击上下箭头调节变量宽度。

小数点后位数（Decimals）：小数点后位数限制数值型、逗号型等变量类型的数值的小数点后的位数。选择变量类型时，可以在“变量类型（Variable Type）”对话框中调整对应的小数点后位数，默认的小数点后位数为 2。用户也可以单击“小数点后位数（Decimals）”该列的单元格，单元格将显示成 2，然后即可在单元格重新输入数值或者单击上下箭头调节小数点后位数。字符型不能限制小数点后位数。

变量标签（Label）：通过变量标签，可以对变量含义进行详细的说明，添加变量名的可视性和分析结果的可读性。例如，在输入调查问卷数据时，由于变量名对于长度有一定的限制，因此变量名就可以用调查问卷的题目号表示，在变量标签中输入完整的问题来进一步解释变量含义。但变量名标签也有长度的限制，其总长度限制在 120 个字符内，并且在统计分析结果显示时，一般不能显示太长的变量标签信息。“变量标签”属性可以省略定义。

数值标签（Value Labels）：通过变量标签，可以对变量数值进行详细的说明。例如，表示性别数据时，用数值“1”表示“男”，数值“2”表示“女”；如果没有数值标签，就很难说明“1”、“2”对应的性别分别是什么。因此，数值标签对于有一定实际含义的数值变量来说是必不可少的。数值标签在明确数值含义的同时，也增强了统计分析结果的可读性。

单击“数值标签（Value Labels）”该列的单元格，单元格右端将出现 … 按钮，单击该按钮，将弹出“值标签（Value Labels）”对话框，如图 2.10 所示。



图 2.10 “值标签（Values）”对话框

在“值标签（Value Labels）”对话框中，可以在“值（Value）”后输入变量中出现的数值，如“1”，然后在标签（Label）中输入对应的数值含义，如“男”，然后单击“添加（Add）”按钮，即可将上述数值标签的定义添加到数值标签的主窗口中。如果需要进行“更改（Change）”或者“删除（Remove）”，可以在主窗口中选择需要更改或移除的数值标签，然后单击对应的按钮。数值标签编辑完成后，单击“确定”按钮，即可返回到数据编辑窗口的变量视图中。

当变量标签和数值标签同时设定完毕后，单击工具栏上的“ （数值标签）”按钮，

数值标签将会取代原来变量的具体数值，用户可以更直接地看出各变量数值所具有的含义。再次单击该按钮，可以取消数值标签的显示，返回至原来的变量数值。

缺失值（Missing Values）：在数据处理过程中，常常会出现空数据或无效数据的情况，即有些数据项漏填了，或者有些数据明显是错误的。这些数据一般都应在数据处理初期时，标示为缺失值数据。

为了在统计分析时，区别对待缺失值和正常值，可以将缺失值填充或更换为特定的标记数据。单击“缺失值（Missing Values）”该列的单元格，单元格右端将出现 ··· 按钮，单击该按钮，将弹出“缺失值（Missing Values）”对话框，如图 2.11 所示。

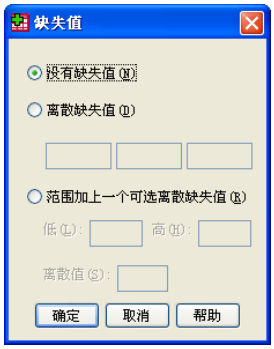


图 2.11 “缺失值（Missing）”对话框

SPSS 中，缺失值的指定方法主要有两种：

（1）字符型或数值型变量：缺失值可以是 1 至 3 个特定的离散值（Discrete missing values）。

（2）数值型变量：缺失值可以在 1 个连续的闭区间内（Range of missing values）或者在 2 个连续的开区间内，并可同时指定 1 个离散值（Range plus one optional discrete missing value）。

SPSS 系统默认的缺失值选项为“没有缺失值（No missing values）”，此时，空数据用一个小圆点表示，但系统默认的缺失值一般只出现在数值型变量数据中。字符型变量中的空格不是系统缺失值。

列宽（Column Width）：变量的列宽（Column Width）和数据的对齐格式（Text Alignment）都属于数据的列格式。显示在数据视图（Date View）中的数据，其变量列宽即是变量类型所对应的变量宽度（Width），默认的列宽为 8。变量的列宽随着变量显示宽度的改变而自动改变。

当变量的显示宽度小于变量的列宽时，需要对变量的列宽进行修改，否则，变量就无法完整地在数据视图中实现。此时，用户可以单击“列宽（Column Width）”该列的单元格，单元格将显示成 8 ，然后即可在单元格中重新输入数值或者单击上下箭头调节变量宽度。

对齐格式（Align）：数据的对齐格式（Align）也属于数据的列格式。显示在数据视图（Date View）中的数据，主要有三种对齐方式：左对齐（Left）、右对齐（Right）、中间对齐（Center）。用户可以单击“对齐（Align）”该列的单元格，然后单击单元格右方的小箭头，如图 2.12 所示，用户即可通过单击，选择需要的对齐方式。



图 2.12 “对齐（Align）”选项

字符型数据默认为左对齐显示，数值型数据默认的对齐格式为右对齐显示。

度量标准（Measure）：统计分析数据的测度方法大致可分为定距度量和非定距度量，对应的数据即为定距数据和非定距数据。定距数据（Scale）是指连续性数据（如人的身高、体重，学生的成绩等）。非定距数据包括顺序尺度（Ordinal）和名义尺度（Nominal）的数据。有内在大小或高低顺序的数据度量即为顺序尺度，如高、中、低职称或者老、中、青三代；不存在内在大小或高低顺序的数据，只包含名义上指代的度量方式即为名义尺度，如男、女性别或者汉、满、回等民族。

系统给出的数据的测度方式包括名义尺度、有序尺度和度量尺度三种。用户可以单击“度量标准（Measure）”该列的单元格，然后单击单元格右方的小箭头，单元格将显示“度量标准（Measure）”选项，如图 2.13 所示。用户即可通过单击，选择需要的测度方式。默认的测度方式为度量（Scale）。



图 2.13 “度量标准（Measure）”选项

定义了变量的各种属性后，单击“Date View”按钮，回到数据视图中，就可以直接在表中输入数据。

除了使用“变量视图（Variable View）”外，SPSS 还提供了对话框定义变量的方法。

选择“数据（Data）”菜单下的“定义变量属性（Define Variable Properties）”命令，SPSS 将弹出“定义变量属性（Define Variable Properties）”对话框，如图 2.14 所示。

SPSS 17.0 for Windows 的“定义变量属性（Define Variable Properties）”对话框没有进行汉化，下面对它进行简要说明。

在“定义变量属性（Define Variable Properties）”对话框中，可以通过单击箭头从左边

变量（Variables）中选择一个或者几个变量进入右框中，进行变量扫描（Variables to Scan）。单击“继续”按钮，即可进入“定义变量属性（Define Variable Properties）”对话框的下一个界面，如图 2.15 所示。

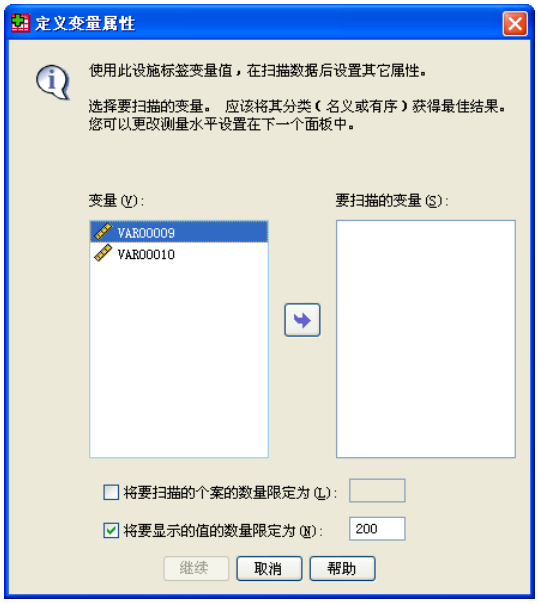


图 2.14　“定义变量属性（Define Variable Properties）”对话框界面 1



图 2.15　“定义变量属性（Define Variable Properties）”对话框界面 2

在“已扫描的变量列表”框中单击想要更改属性的变量，右边的变量属性列表中将显示该变量的所有当前属性，用户可以单击对应窗口进行数值或方式的修改，各属性的具体意义

与上述变量视图（Variable View）中的属性意义相同，此处不再赘述。同时，SPSS 的“定义变量属性（Define Variable Properties）”对话框提供了复制属性的功能，用户可以直接从其他变量复制属性（From Another Variable），也可以将当前变量的属性复制给其他变量（To Other Variables）。修改完成后，单击“确定”按钮，即可完成修改，并保存变量属性。

2.2.3 插入或删除数据

插入或删除数据包括插入或删除个案（Case）和变量（Variable）。在数据编辑过程中，常常需要在数据编辑窗口的某个个案（变量）之前或之后插入一个新个案（变量），或者删除某些不再有用的个案（变量）。

插入一个新个案的命令是标识异常个案（Insert Cases）。在数据窗口主菜单上单击“数据（Date）”下的“标识异常个案（Insert Cases）”命令，如图 2.16 所示，即可在光标所在位置的前一行插入一行新的个案，随即可以输入新的观测数据。插入一个新变量的命令是插入变量（Insert Variable）。在数据窗口主菜单上单击“编辑（Edit）”下的“插入变量（Insert Variable）”命令，即可在光标所在位置的前一列插入一列新的变量，变量名和变量属性可以在变量视图（Variable View）窗口中进行定义。



图 2.16 “数据（Date）”菜单

用户也可以直接单击工具栏上的 和 ，分别标识异常个案和变量。

删除个案（Delete Cases）时首先要选中将要删除的个案数据行，待删除的个案数据将反白显示。单击鼠标右键，从弹出的快捷菜单中选择“清除（Clear）”命令，如图 2.17 所示。待删除的个案数据将被删除，其后的数据将自动依次前移。

删除变量（Delete Variables）时首先要选中将要删除的变量数据列，待删除的变量数

据将反白显示。单击鼠标右键，从弹出的快捷菜单中选择“清除（Clear）”命令，如图 2.18 所示。待删除的变量数据将被删除，其右面的数据将自动依次左移。



图 2.17　右击个案后弹出的快捷菜单



图 2.18　右击变量后弹出的快捷菜单

2.2.4　数据的排序

数据编辑窗口中个案数据的顺序是由数据输入时的先后顺序决定的，一般是随机的。但在实际数据处理过程中，有时用户希望将个案数据按照某种顺序进行排列，以便进行观察和分类，此时就要用到 SPSS 中的排序命令。

SPSS 的排序个案就是根据用户指定的一个或多个变量的变量值的大小，按照升序或降序将个案数据进行重新排列。这里用户所指定的变量称为排序变量。当排序变量只有一个时的排序称为单值排序，即：按照排序变量取值的大小次序对个案数据重新排列后显示。当排序变量有多个时的排序称为多重排序。多重排序中，排序变量根据用户指定的先后次序，依次分为第一排序变量、第二排序变量、第三排序变量……即在进行多重排序时，个案数据首先按照第一排序变量的大小次序进行排列，对于具有相同第一排序变量数值的个案，接着按照第二排序变量的大小次序进行排列，再对那些具有相同第二排序变量数值的个案，按照第三排序变量的大小次序进行排列，依次类推。

SPSS 中，给个案数据进行排序的命令是“排序个案（Sort Cases）”。具体操作步骤如下。

01 在数据编辑窗口中单击“数据（Date）”，选择“排序个案（Sort Cases）”，打开“排序个案（Sort Cases）”对话框，如图 2.19 所示。



图 2.19　“排序个案（Sort Cases）”对话框

02 在“排序个案（Sort Cases）”对话框中，从对话框左侧的源变量列表中，按照排序变量重要程度的先后次序，选择一个或几个排序变量，单击按钮，加入“排序依据（Sort by）”框中。

03 在“排序顺序（Sort Order）”栏中选择每个排序变量的排序顺序。

升序（Ascending）：个案按照选定的排序变量值由小到大升序排列；

降序（Descending）：个案按照选定的排序变量值由大到小降序排列。

注意在排序的时候，个案整行的数值将一同变到新位置，而不是只有排序变量那一列的数值发生变动。但数据编辑窗口中左边框的序号并不随着个案的变动而变动，因此，为了保留原始数据的信息，建议用户自行定义一个变量来记录个案个体的序号。

04 在“排序个案（Sort Cases）”对话框中，单击“确定”按钮，即可在数据编辑窗口得到排序个案后的新数据文件。

2.2.5 选择个案

在数据处理过程中，有时需要从数据文件中选取一部分个案，将筛选出的个案进行统计分析的操作、处理和分析，而未被选取到的个案将不参与之后的操作。选择个案的操作步骤如下。

01 在数据编辑窗口中，选择“数据（Data）”下的“选择个案（Select Case）”命令，打开“选择个案（Select Case）”对话框，如图 2.20 所示。

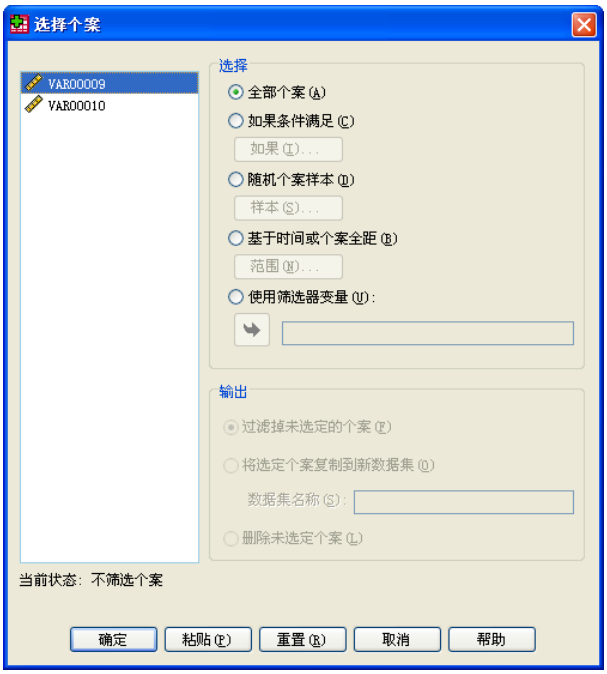


图 2.20 “选择个案（Select Case）”对话框

02 在“选择（Select）”栏中选择挑选数据子集的方式（单选项）。

全部个案（All cases）：所有的个案都被选择，作为数据处理的对象。

如果条件满足（If condition is satisfied）：根据用户指定的条件，即用户给出数据需要满足的条件表达式，SPSS 对数据编辑窗口中的所有个案进行判断，筛选出符合条件的数据个案。单击“如果（If）”按钮，打开“选择个案：If”对话框，如图 2.21 所示。用户需要先从对话框左侧的源变量列表中选择一个或几个条件变量，然后利用对话框中的数字、数学符号和函数定义条件，条件表达式将在对话框右上方的窗口中显示。条件设定完毕，单击“继续”按钮，即可返回“选择个案（Select Case）”对话框。



图 2.21 “选择个案：If”对话框

随机个案样本（Random Sample of cases）：对数据编辑窗口中的所有观察量进行随机抽样。单击“样本（Sample）”按钮，打开“选择个案：随机样本（Select Case：Random sample）”对话框，如图 2.22 所示。在“样本尺寸（Sample Size）”栏中，SPSS 提供了两种随机样本的选择方式。一种是“大约（Approximately）”，即用户输入抽样比例后由系统按照该百分比进行随机抽样；但这种情况下，由于 SPSS 在实际计算中存在的近似性和在个案选取处理方面的技术要求，可能会出现微小的偏差，但对于数据的后续分析不会产生影响。SPSS 提供的另一种抽样方式是“精确（Exactly）”，即要求用户输入从第几个观察值开始抽取多少个数的样本数据，之后，SPSS 即可在设定的范围内精确地抽取出随机样本个案。样本选择方式设定完毕后，单击“继续”按钮，即可返回“选择个案（Select Case）”对话框。

基于时间或个案全距（Based on time or case range）：按照时间或者个案序号的顺序抽取某一区域内的全部个案。单击“范围（Range）”按钮，打开“选择个案：范围（Select Case：Range）”对话框，如图 2.23 所示。用户需要自行定义开始的观察值（First Case）到“最后观测”的观察值（Last Case），SPSS 将抽取出该范围内的所有个案。这种个案选取的方式

比较适用于时间序列的数据分析。范围设定完毕后，单击“继续”按钮，即可返回“选择个案（Select Case）”对话框。

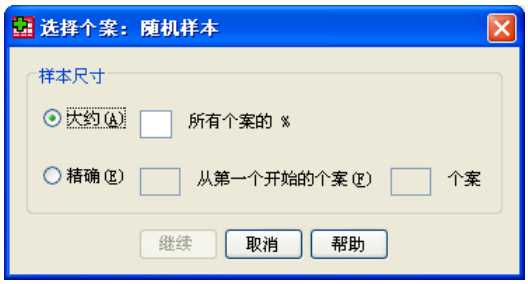


图 2.22 “选择个案：随机样本（Select Case：Random sample）”对话框

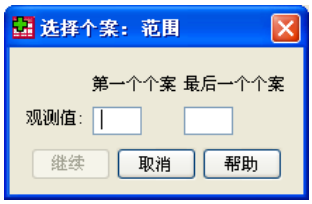


图 2.23 “选择个案：范围（Select Case：Range）”对话框

使用筛选器变量（Use filter variable）：用户指定一个已存在的变量作为个案选取的标准，这个作为筛选标准的变量称为过滤变量。SPSS 系统自动在数据管理器中将过滤变量值为 0 或系统缺失值的个案标上删除记号，标有删除记号的个案将被剔除，不参与后续分析。

需要说明的是，如果采用如果条件满足（If condition is satisfied）、随机样本个案（Random Sample of cases）、按时间或个案全距（Based on time or case range）来选取个案，SPSS 将在数据编辑区中自动生成一个名为“filter_$”的新变量。新变量的数值为 0，表示本个案未被选中；新变量的数值为 1，表示本个案已被选中。同时，SPSS 自动为该变量添加变量名标签，以方便用户直接了解本次个案筛选的方式。

03 在“输出（Output）”栏中指定未选中的个案的处理方式（单选项）。

过滤未选定的个案（Filter out unselected cases）：将未被选中的个案的个案号码前标记为“/”，以表示与已选中个案的区别。

将选定个案复制到新数据集（Copy selected cases to a new dataset）：单击该按钮后，在窗口中输入新的数据文件的名字，SPSS 将自动打开一个以新的数据文件的名字命名的数据编辑窗口，窗口中包括所有已被选中的数据。

删除未选定个案（Delete unselected cases）：将未被选中的个案直接从数据文件中删除，此处理方法风险较大，不可恢复。

04 在“选择个案（Select Case）”主对话框中，单击“确定”按钮，即可在数据编辑窗口根据不同的输出方式得到选择个案的结果数据文件。

2.2.6 转置数据

数据文件的转置就是将数据编辑窗口中数据的行列互换，即将个案转换为变量，将变量转换为个案，然后重新显示在数据编辑窗口中。转置数据的操作步骤如下。

01 选择“数据（Data）”菜单下的“转置（Transpose）”按钮，打开“转置（Transpose）”

对话框，如图 2.24 所示。



图 2.24 “转置（Transpose）”对话框

02 在“转置（Transpose）”对话框中，用户可以从对话框左侧的源变量列表中，指定需要转置的变量，单击 按钮加入“变量（Variables）”列表框中。在“变量（Variables）”列表框中，一般应选择除名称变量外的所有其他变量，如果有变量未选入，则转置时会被自动丢弃。“名称变量（Name Variable）”用于指定原数据文件转置后表示变量名的变量，这要求该变量的每个个案数值都必须是唯一的，转置后的数据文件中的各个变量将以名称变量的各个观测值命名。但名称变量不是必需的，如果没有名称变量，系统会默认将新数据变量按 VAR00001、VAR00002……的顺序命名。

03 单击“确定”按钮后，数据编辑窗口中将显示数据文件转置后的数据，SPSS 系统将自动生成一个名为 Case-lbl 的新变量，用于保存原数据文件的各个变量名。

2.2.7　合并数据文件

在进行数据输入时，经常遇到这样的情况：如果数据的总体数量较大，需要把一份大的数据分成几个小部分，然后再分别由多人输入，以节省数据输入时间。此时，就会出现同一份数据的不同部分储存在不同数据文件中的情况。因此，在进行统计分析前，首先要将这若干份数据合并成原始的大的数据文件。

在 SPSS 中，数据文件的合并是指将一个已储存在磁盘上的 SPSS 数据文件与当前 SPSS 数据编辑窗口中的数据文件进行合并，并在当前数据编辑窗口中显示合并后数据的过程。因此，实行两个 SPSS 数据文件的合并时，首先应打开当前数据编辑窗口中的一个数据文件，然后再实行与另一数据文件的合并。

SPSS 数据文件的合并方式分为两种：纵向合并和横向合并。

从外部数据文件中将记录数添加（即个案个数）到当前数据文件中，称为纵向合并。通过纵向合并，能够实现两个数据文件中数据的首尾对接，合并后的数据将重新显示在当前数据编辑窗口中。纵向合并的实质即为将两个数据文件的变量列，按照相同的变量名，一一对应进行首尾连接。纵向合并的操作步骤如下。

01 选择“数据（Data）”菜单下的“合并文件（Merge Files）”命令，单击“添加个案（Add Cases）”按钮，打开“将个案添加到…（Add Cases to）”对话框，如图 2.25 所示。

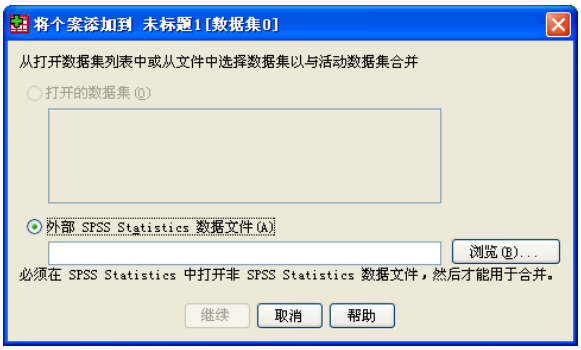


图 2.25 “将个案添加到…（Add Cases to）”对话框

02 选择纵向合并数据文件的方式，SPSS 提供了两种方式供用户选择：

一种是直接从已打开的数据文件中合并，这样就无须将欲合并的数据文件的数据编辑窗口关闭。用户选中“已打开数据文件（An open dataset）”单选框，然后在下方的窗口中单击已打开数据文件的文件名即可。

另一种是从外部的 SPSS 数据文件中合并，即选中“外部 SPSS 数据文件（An external SPSS data file）”后，单击“浏览（Browse）”按钮，SPSS 将弹出“添加个案：读取文件（Add Cases：Read File）”对话框，如图 2.26 所示。“添加个案：读取文件（Add Cases：Read File）”对话框与打开数据文件的对话框相同。用户选中需要合并的数据文件，并单击“打开”按钮，即可返回“将个案添加到…（Add Cases）”对话框中。

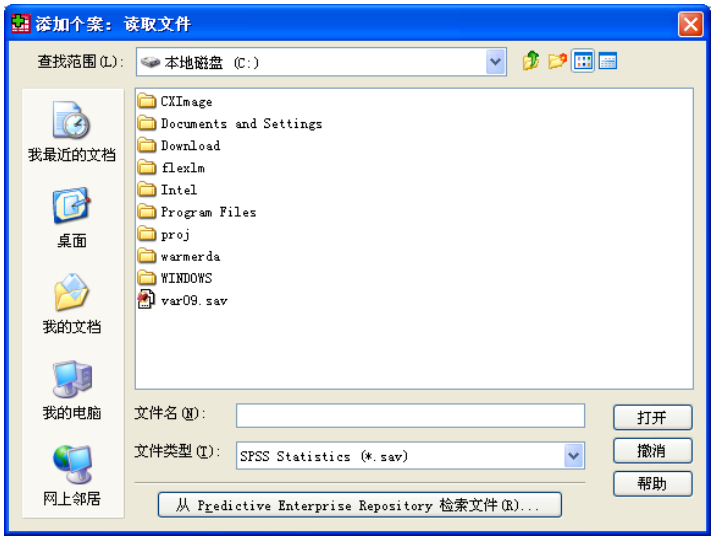


图 2.26 “添加个案：读取文件（Add Cases：Read File）”对话框

03 选定需要合并的数据文件后，单击“继续”按钮，SPSS 将弹出“添加个案从…（Add Cases from）”对话框，如图 2.27 所示。

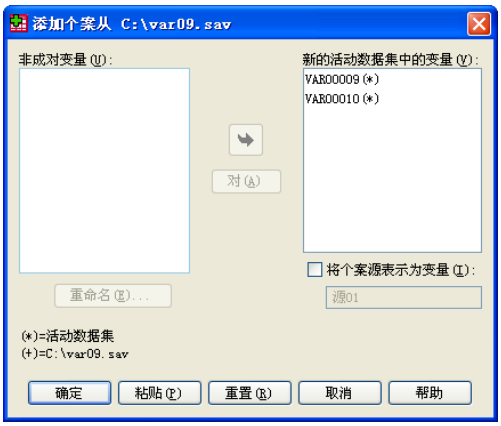


图 2.27 “添加个案从…（Add Cases from）”对话框

在“添加个案（Add Cases）”对话框中，左侧的变量框显示的是新、老数据文件中不配对的变量名（Unpaired Variables），右侧变量框显示的是 SPSS 自动配对的变量名，即新工作数据文件中的变量（Variables in New Active Dataset）。

在非成对变量（Unpaired Variables）框中，变量名后面会出现“*”号或“+”号。“*”号表示该变量名是当前数据编辑窗口中的变量，“+”号表示该变量名是用户指定合并的磁盘文件中的变量。由于变量名无法进行配对，SPSS 将其默认为不具有相同的数据含义。此种情况下，用户可以进行手工配对变量：选中非成对变量（Unpaired Variables）框中的欲配对的变量名，单击“对（Pair）”按钮，变量将进入“新工作数据文件中的变量（Variables in New Active Dataset）”对话框中，表示它们具有相同的数据含义。用户也可以选择某个变量名，单击“更名（Rename）”按钮修改变量名称后，再进行指定配对。如果有的变量在另一文件中不具有任何数据或意义，但用户仍想将两个文件合并，用户可以直接单击 按钮，不进行任何配对，强行进入“新工作数据文件中的变量（Variables in New Active Dataset）”对话框中。

当然，如果用户不接受匹配变量名的自动配对，也可以选中“新工作数据文件中的变量（Variables in New Active Dataset）”，单击 按钮取消自动配对。

右下方的“将个案源表示为变量（Indicate case source as variable）”复选框用于定义一个新变量以区分个案来自原数据编辑窗口还是添加的数据文件。选中该选项，合并后的数据文件中将自动出现一列变量名为 source01 的变量，变量的取值为 0 或 1。0 表示该个案来自原数据编辑窗口，1 表示该个案来自添加的数据文件。

04 选择完毕后，单击“确定”按钮，即可实现纵向合并。

从外部数据文件中将变量添加到当前数据文件中，称为横向合并。通过横向合并，能够实现将两个数据文件中数据的左右对接，合并后的数据将重新显示在当前数据编辑窗口

中。横向合并的实质即为将两个数据文件的数据，按照相同的个案名或者序号一一对应地进行左右连接。

实现 SPSS 数据文件的横向合并应满足以下三个条件：

（1）两个数据文件必须至少有一个变量名相同的公共变量，如姓名、序号等，这个公共变量将作为两个数据文件横向对应拼接的依据，因此又被称为关键变量（Key Variables）；

（2）两个数据文件必须事先按照公共变量进行升序排序或者多重升序排序；

（3）为了方便 SPSS 数据文件的横向合并，在不同数据文件中，数据含义不同的列、变量名也应以不同的名称命名。

横向合并的操作步骤如下。

01 选择“数据（Data）”菜单下的“合并文件（Merge Files）”命令，单击“添加变量（Add Variables）”按钮，打开“将变量添加到…（Add Variables to）”对话框，如图 2.28 所示。

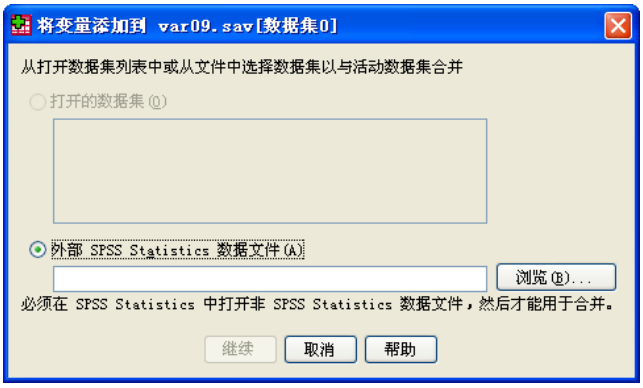


图 2.28 “将变量添加到…（Add Variables to）”对话框

02 选择横向合并数据文件的方式。同样地，SPSS 提供了两种方式供用户选择。

一种是直接从已打开的数据文件中合并，这样就无须将欲合并的数据文件的数据编辑窗口关闭。用户选中“打开的数据集（An open dataset）”单选框，然后在下方的窗口中单击已打开数据文件的文件名即可。

另一种是从外部的 SPSS 数据文件中合并，即选中“外部 SPSS Statistics 数据文件（An external SPSS data file）”后，单击“浏览（Browse）”按钮，SPSS 将弹出“添加变量：读取文件（Add Variables：Read File）”对话框，如图 2.29 所示。“添加变量：读取文件（Add Variables：Read File）”对话框与打开数据文件的对话框相同。用户选中需要合并的数据文件，并单击“打开”按钮，即可返回“将变量添加到…（Add Variables to）”对话框中。

03 选定需要合并的数据文件后，单击“继续”按钮，SPSS 将弹出“添加变量从…（Add Variables from）”对话框，如图 2.30 所示。

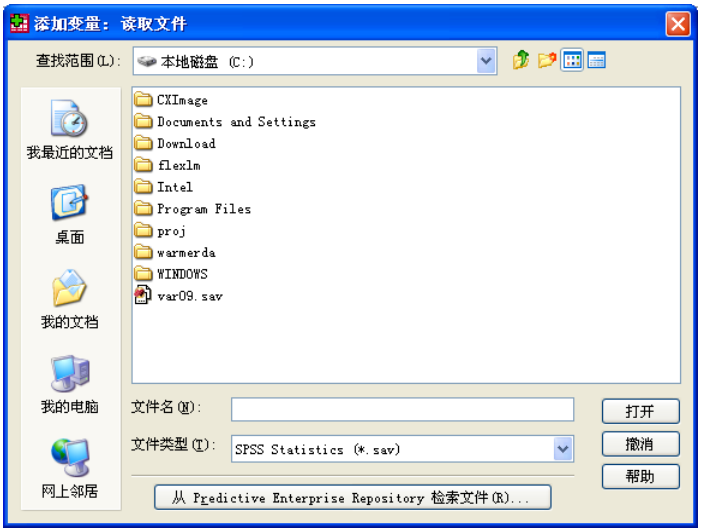


图 2.29　“添加变量：读取文件（Add Variables：Read File）”对话框

在“添加变量从…（Add Variables from）”对话框中，两个待合并的数据文件的所有变量名将出现在“新的活动数据集（New Active Dataset）”列表框中，共有变量名将出现在“已排除的变量（Excluded Variables）”列表框中。

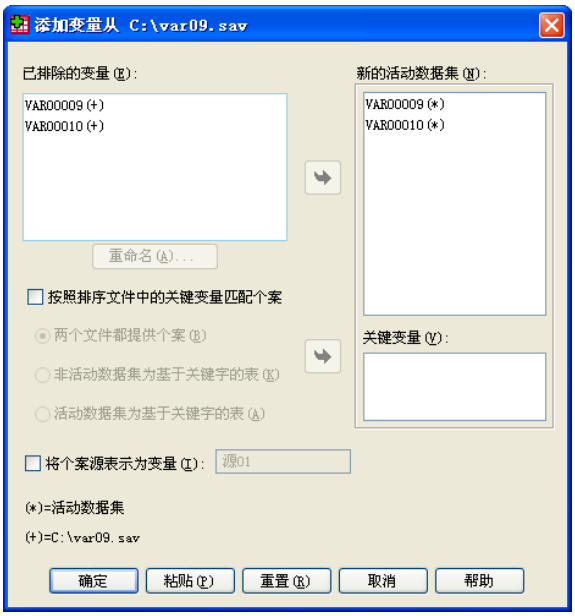


图 2.30　“添加变量从…（Add Variables from）”对话框

在“新的活动数据集（New Active Dataset）”列表框中，所有的变量名后面会出现“ * ”号或“ + ”号。“ * ”号表示该变量名是当前数据编辑窗口中的变量，“ + ”号表示该变量

名是用户指定合并的磁盘文件中的变量。SPSS 默认这些变量以原变量名称成为合并后新工作数据文件中的变量。如果用户不需要全部的原变量，可以进行手工剔除变量：选中“新工作数据文件（New Active Dataset）”列表框中的欲剔除的变量名，单击 按钮，变量将进入“已排除的变量（Excluded Variables）”列表框中。如果用户不满意原变量的名称，也可以选择某个变量名进入“已排除的变量（Excluded Variables）”列表框后，单击“重命名（Rename）”按钮修改变量名称，再单击 按钮，将原变量以新的名字重新选回“新的活动数据集（New Active Dataset）”列表框中。

选中“按照排序文件中的关键变量匹配个案（Match cases on key variables in sorted files）”选项，用户可以在“已排除的变量（Excluded Variables）”列表框中选择一个或多个变量作为关键变量，单击 按钮添加到“关键变量（Key Variables）”列表框中。SPSS 将根据关键变量中相同的个案值一一配对进行左右连接。

同时，选中“按照排序文件中的关键变量匹配个案（Match cases on key variables in sorted files）”选项后，SPSS 提供了以下三种指定合并后数据文件中的数据来源方式。

（1）两个文件都提供个案（Both files provide cases）：SPSS 的默认选项，直到合并后数据文件中的数据由两个数据文件中的个案共同组成，即对于关键变量中数据不同的个案，SPSS 也将其合并到新的数据文件中，缺失的数据以缺失值的形式显示。

（2）非活动数据集为基于关键字的表（Non-active dataset is keyed table）：在外部数据文件中的数据基础之上，合并当前数据编辑窗口中的数据变量，即对于关键变量中数据不同的个案，SPSS 仅添加与外部数据文件中的数据的关键变量数值相同的个案，其他变量予以放弃。

（3）活动数据集为基于关键字的表（Active dataset is keyed table）：与外部文件为关键表相反，在当前数据编辑窗口中的数据的基础之上，合并外部数据文件中的数据变量，即对于关键变量中数据不同的个案，SPSS 仅添加与当前数据编辑窗口中的数据的关键变量数值相同的个案，其他变量予以放弃。最下方的“将个案源表示为变量（Indicate case source as variable）”复选框用于定义一个新变量，以区分个案来自原数据编辑窗口还是添加的数据文件。选中该选项，合并后的数据文件中将自动出现一列变量名为 source01 的变量，变量的取值为 0 或 1。0 表示该个案来自原数据编辑窗口，1 表示该个案来自添加的数据文件。

04 选择完毕后，单击“确定”按钮，即可实现横向合并。

2.2.8 数据的分类汇总

所谓数据分类汇总，就是按照用户指定的分类变量对个案进行分组，并对每组个案的各变量值计算指定的描述统计量（如求和、平均值等）。通过分类汇总，用户可以针对不同的组别了解每组的大致情况，并可以做出比较，例如分别统计某校高一男女生的语文平均成绩。实现数据分类汇总的方式如下。

1．选择“数据（Data）”菜单下的“分类汇总（Aggregate）”命令，打开“汇总数据

（Aggregate Data）”对话框，如图 2.31 所示。

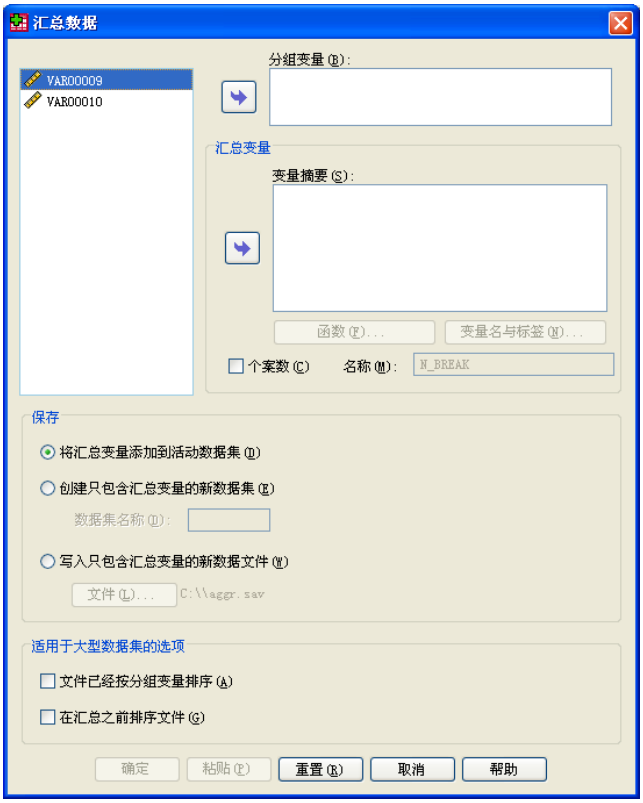


图 2.31　“汇总数据（Aggregate Data）”对话框

在“汇总数据（Aggregate Data）”对话框中，左边的变量列表为源变量列表，通过单击按钮，可选择一个或者几个变量进入右边的“中断变量（Break Variable(s)）”列表框中，作为分组变量。

同时，用户还需要单击按钮，选择变量进入“汇总变量（Aggregated Variables）”下的“变量摘要（Summaries of Variable(s)）”列表框中，作为被汇总的统计变量。“汇总变量（Aggregated Variables）”栏中的其他按钮意义如下。

函数（Function）：用于定义汇总函数，计算指定的描述统计量。单击“函数（Function）”按钮，弹出“汇总数据：汇总函数（Aggregate Data：Aggregate Function）”对话框，如图 2.32 所示。SPSS 提供了三组函数，但每个汇总变量只能选择一个描述统计量进行汇总。一般来说，第一组函数（摘要统计、特定值、个案数）最为常用，用户直接选中描述统计量对应的函数按钮即可。但用户如果选择第二组函数（百分比）或第三组函数（分数），用户除需选中选项外，还需在小窗口中输入需要统计的值、下限或上限的数值。进行汇总方式的选择后，单击“继续”按钮，即可返回“汇总数据（Aggregate Data）”对话框。

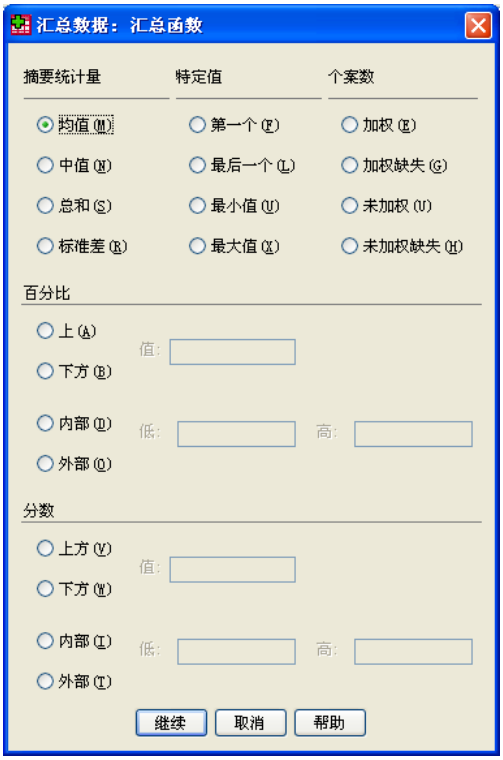


图 2.32 “汇总数据：汇总函数（Aggregate Data：Aggregate Function）”对话框

名称和标签（Name & Label）：用于定义新产生的汇总变量的名称和标签。单击“名称和标签（Name & Label）”按钮，弹出“汇总数据：变量名称和标签（Aggregate Data：Variable Name and Label）”对话框，如图 2.33 所示。用户可以在对话框内输入汇总变量的名称和标签，定义完毕后，单击“继续”按钮，即可返回“汇总数据（Aggregate Data）”对话框。



图 2.33 “汇总数据：变量名称和标签（Aggregate Data：Variable Name and Label）”对话框

2．选择保存分类汇总的方式。进行分类汇总后，分类汇总的结果既可以存入新数据文件，也可以替换当前数据文件。“汇总数据（Aggregate Data）”对话框的“保存”栏就提供了下列三种保存分类汇总结果的方式。

- 将汇总变量添加到活动数据集（Add aggregated variables to active dataset）：SPSS 将自动定义一个新变量，用以储存分类汇总的结果，新变量将自动添加到当前数据文件中；
- 创建只包含汇总变量的新数据集（Create a new dataset containing only the aggregated variables）：选中该选项，可以单击“文件（File）”按钮，指定汇总文件的保存路径和文件名，SPSS 将创建只包含汇总结果的新数据文件；
- 写入只包含汇总变量的新数据文件（Write a new data file containing only the aggregated variables）：选中该选项，分类汇总的结果将覆盖当前数据编辑窗口中的数据，原数据将丢失。

3．针对个案个数较多的数据文件，SPSS 提供了“适用于大型数据集选项（Options for Vary Large Datasets）”栏（复选项），方便此类数据文件的分类汇总。

- 文件已经按分组变量排序（File is already sorted on break variable(s)）：一般情况下，特别是个案个数较多的数据文件，在进行分类汇总前，用户需要将个案数据根据中断变量（即分组变量）进行排序；否则，分类汇总无法进行。如果用户已经手动完成排序，即可选择该选项，SPSS 将忽略排序的步骤，自动开始分类汇总。
- 在汇总之前排序文件（Sort file before aggregating）：如果用户选择“分类汇总”前，并未对数据文件中的数据按照中断变量进行排序，也没有关系。SPSS 提供了汇总之前先对文件进行排序的功能。选择该选项，SPSS 将先对数据进行排序，然后再进行分类汇总。

4．选择完毕后，单击“确定”按钮，即可实现分类汇总。

2.2.9　数据菜单的其他功能

“加权个案（Weight Cases）”对话框

在实际工作中经常需要计算加权平均数。在计算过程中，利用权数变量对数据进行加权处理，就需要用到“加权个案（Weight Cases）”对话框。选择“数据（Data）”菜单下的“加权个案（Weight Cases）”命令，SPSS 将弹出“加权个案（Weight Cases）”对话框，如图 2.34 所示。

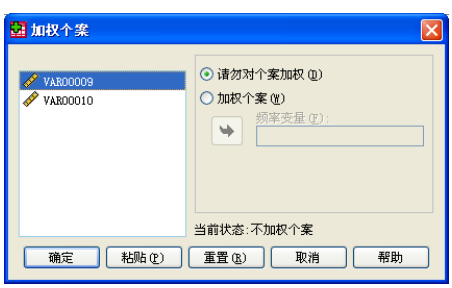


图 2.34　“加权个案（Weight Cases）”对话框

对话框界面上有两个单选按钮：

- 请勿对个案加权（Do not weight cases）：SPSS 的默认选项，不对数据进行加权处理；
- 加权个案（Weight cases by）：选择该选项，需要用户从对话框左侧的变量列表中选中一个频数变量（Frequency Variable）作为权重变量，SPSS 将自动进行加权处理。

“定义日期（Define dates）”对话框

在对数据进行统计分析的过程中，特别是进行时间序列分析时，常常需要用到时间变量。如果每次定义时间变量都是以手动输入日期的方式，那显然过于麻烦。因此，SPSS 提供了一个“定义日期（Define dates）”的功能，以简化定义时间变量的过程。选择“数据（Data）”菜单下的“定义日期（Define dates）”命令，SPSS 将弹出“定义日期（Define dates）”对话框，如图 2.35 所示。

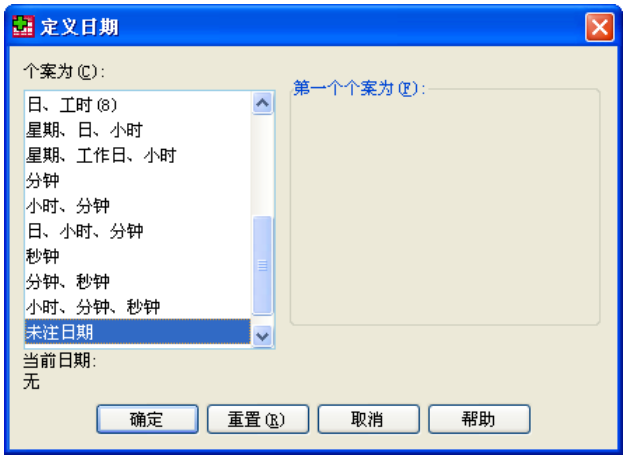


图 2.35 “定义日期（Define dates）”对话框

用户可以在对话框中选择所定义的日期格式，如年（Years）、年和季度（Years, quarters），然后在右端的小窗口中输入该日期格式所对应的起始时间等值。单击“确定”按钮，即可完成操作。SPSS 将在当前数据编辑窗口中自动生成一列或几列（根据日期格式的不同）的时间变量。

2.3 数据加工

2.3.1 数据转换

数据转换就是在原有数据文件的基础之上，根据用户所给出的 SPSS 算术表达式和函数，对某个变量的所有个案进行计算，并将计算结果保存至数据编辑窗口中的一个指定变量中。这个指定的变量可以是生成的一个新变量，也可以替代原来已经存在的变量。

在 SPSS 中，常常需要用到数据转换，进而生成新变量。例如，在给所有学生的数学

成绩都加上 20 分时，或者需要计算所有学生的数学和物理的总成绩时，都需要使用数据转换。SPSS 提供了“计算变量（Compute Variable）”命令，来满足用户的数据转换的需求。数据计算转换的操作步骤如下。

01 单击“转换（Transform）”菜单下的“计算变量（Compute Variable）”命令，如图 2.36 所示。SPSS 将弹出“计算变量（Compute Variable）”对话框，如图 2.37 所示。



图 2.36 “转换（Transform）”菜单



图 2.37 “计算变量（Compute Variable）”对话框

02 在“计算变量（Compute Variable）”对话框左上方的“目标变量（Target Variable）”栏中，需要输入即将生成的新变量的名称。输入变量名后，下方的“类型与标签（Type & Label）”按钮则呈现为可选状态，单击“类型与标签（Type & Label）”按纽，弹出“计算变量：类型和标签（Compute Variable: Type & Label）”对话框，如图 2.38 所示。在该对话框中，用户可以定义新变量的标签及数据类型。

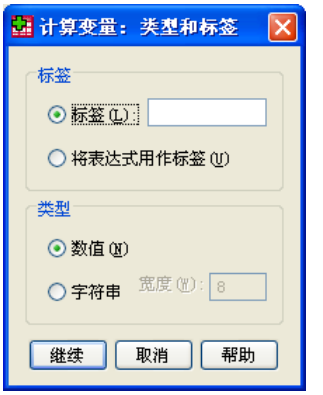


图 2.38 “计算变量：类型和标签（Compute Variable: Type & Label）”对话框

“计算变量：类型和标签（Compute Variable: Type & Label）”对话框的“标签（Label）”栏中，提供了两种定义新变量标签的方法：

- 标签（Label）：用户需要在后面的小窗口中输入字符定义新变量标签；
- 将表达式用作标签（Use expression as label）：使用用户定义的运算表达式作为标签。

同时，该对话框的“类型（Type）”栏中只提供了两种变量类型——数值（Numeric）和字符串（String）。如果用户需要将新变量设定为其他的变量类型，可以返回至数据编辑窗口后在变量视图（Variable View）中进行修改。

03 “计算变量（Compute Variable）”对话框的左下栏是源变量列表，列表中给出了数据文件中所有可用的变量。通过单击可以选择所需的变量加入右上方的“数字表达式（Numeric Expression）”栏中，作为运算表达式的一部分。SPSS 的运算表达式是由常数（1~10）、SPSS 变量名、SPSS 的算术运算符（+、–等）、SPSS 的常用函数以及圆括号等组成的式子。运算表达式中所需要的常数、算术运算符等可以从键盘直接输入，也可以通过单击“数字表达式（Numeric Expression）”栏下方的数字和运算符的小键盘中选取。SPSS 的常用函数也可以从“数字表达式（Numeric Expression）”栏下方的“函数组（Function Group）”列表中选择，选择某一函数组后，“函数组（Function Group）”列表框下的“函数和特殊变量（Functions and Special Variables）”将显示该函数组所包含的所有常用函数，单击鼠标左键可以直接选取。同时，“函数和特殊变量（Functions and Special Variables）”左侧的空白窗口中将显示所选中的常用函数的使用说明。这些常用函数（见附录 1）和其

他软件语言中的函数名称类似，在列表框中按照字母顺序排列，用鼠标选中某个函数后，用“函数组（Function Group）”列表左面的 加入运算表达式中。

04 在“计算变量（Compute Variable）”对话框的左下方还有一个“如果（If）”按钮，单击该按钮，SPSS 将弹出“计算变量：If 个案（Compute Variable：If Cases）”对话框。如图 2.39 所示。在该对话框中，用户可以设定个案的条件，只有一部分满足某种条件的个案才能进行运算，而不满足条件的个案，其新变量值将显示为缺失。设定个案的条件后，单击“继续”按钮，即可返回“计算变量（Compute Variable）”对话框。

05 在“计算变量（Compute Variable）”对话框中定义新变量和运算表达式后，单击“确定”按钮，即可生成新变量，完成数据转换。



图 2.39　“计算变量：If 个案（Compute Variable：If Cases）”对话框

2.3.2　数据手动分组（编码）

在实际工作中，人们经常需要将统计数据进行分组，以进一步整理数据。对变量数据的分组（也可以称为编码），是指给每个变量观测值重新赋予一个码来描述它们的某些属性，并把码数相同的分为一组。例如，对年龄进行分组时，可以将 19 岁及以下的年龄赋予编码 1，20～29 岁的年龄赋予编码 2，30～39 岁的年龄赋予编码 3，依次类推；那么年龄即被分为了三个不同的组。从某种程度上来讲，编码可以看做是分组的过程，一个组对应一个码号。

可见，数据分组（编码）主要分为以下三个步骤：

01 指定分组变量，即指定将哪个变量的个案进行分组；

02 定义分组变量的分组区间，并给每个区间赋予对应的码；

03 指定存放分组结果的标志变量，并标志出每个组的组号（即码）。

通过编码和分组，用户可以一目了然地了解数据的大体情况。

SPSS 提供了两种数据分组的方式：一种是数据自动分组（参见 2.3.3 节），另一种是数据手动分组。数据手动分组还分为重新编码（分组）到相同变量（Recode into Same Variables）和重新编码（分组）到不同变量（Recode into Different Variables）两类。下面将分别进行说明。

重新编码（分组）到相同变量（Recode into Same Variables）是将数据编码后，新的码值直接替换原来的变量。为了避免数据丢失，尽量不要轻易选择该项。重新编码（分组）到相同变量的操作步骤如下。

01 单击“转换（Transform）”菜单下的“重新编码为相同变量（Recode into Same Variables）”选项，SPSS 将弹出“重新编码到相同的变量中（Recode into Same Variables）”对话框，如图 2.40 所示。



图 2.40 “重新编码到相同的变量中（Recode into Same Variables）”对话框

在“重新编码到相同的变量中（Recode into Same Variables）”对话框中，用户可以从对话框左侧的源变量列表中，指定需要进行分组的变量，单击 按钮加入“变量（Variables）”列表框中。

02 与“计算变量（Compute Variable）”对话框相同，“重新编码到相同的变量中（Recode into Same Variables）”对话框下方有一个“如果（If）”按钮，单击该按钮，SPSS 将弹出“重新编码成相同变量：If 个案（Recode into Same Variables：If Cases）”对话框，如图 2.41 所示。在该对话框中，用户可以设定个案的条件，只有一部分满足某种条件的个案才能进行分组。设定个案的条件后，单击“继续”按钮，即可返回“重新编码到相同的变量中（Recode into Same Variables）”对话框。



图 2.41　“重新编码成相同变量：If 个案（Recode into Same Variables：If Cases）”对话框

03 选择需要重新分组（编码）的变量后，单击“变量（Variables）”列表框下方的“旧值和新值（Old and New Values）”按钮，打开“重新编码成相同变量：旧值和新值（Recode into Same Variables：Old and New Values）”对话框，如图 2.42 所示。

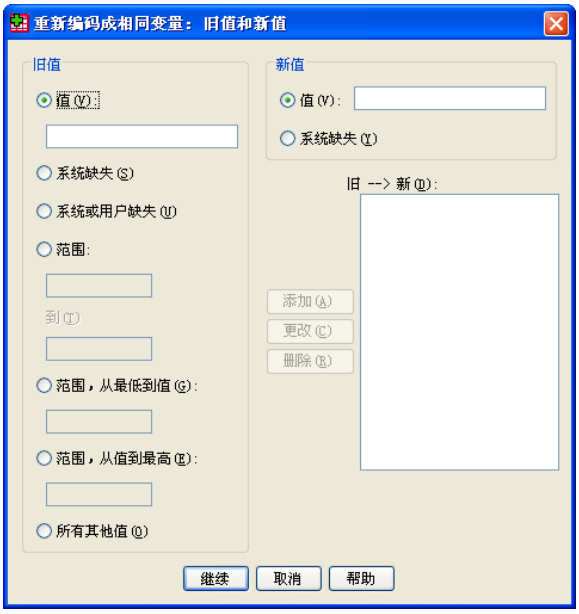


图 2.42　“重新编码到相同变量：旧数值和新数值（Recode into Same Variables：Old and New Values）”对话框

对话框左侧为“旧值（Old Value）”栏，有七个选项，用来确定原变量的连续取值区间（或单个变量观测值），它们将被赋予一个具有相同意义的新码值。右上方的“新值（New

Value）”栏中可以填入新的将替代原数据的码值。一组对应的新旧值填写完毕后，“添加（Add）”按钮将被激活。单击“添加（Add）”按钮，旧的变量区间（值）以及对应的新的码值将被添加到“旧值→新值（Old→New）”列表框中。重复以上步骤，直到所有的旧的变量区间（值）和对应的新的码值都完成输入为止。单击“继续”按钮，即可返回“重新编码到相同的变量中（Recode into Same Variables）”对话框中。

04 在“重新编码到相同的变量中（Recode into Same Variables）”主对话框中单击“确定”按钮，即可执行命令，在数据编辑窗口中得到需要的分组赋值变量。原变量已被新的分组赋值变量所覆盖。

重新编码（分组）为不同变量（Recode into Different Variables）是将数据编码后新的码值存到一个新变量中，不覆盖源变量，因而，不会影响原始数据的存放。重新编码（分组）为不同变量的操作步骤如下。

01 单击“转换（Transform）”菜单下的“重新编码为不同变量（Recode into Different Variables）”命令，SPSS 将弹出“重新编码为其他变量（Recode into Different Variables）”对话框，如图 2.43 所示。

“重新编码为其他变量（Recode into Different Variables）”对话框与“重新编码成相同变量（Recode into Same Variables）”对话框相似，用户可以从对话框左侧的源变量列表中，指定需要进行分组的变量，单击 按钮加入“输入变量→输出变量（Input Variable→Output Variable）”列表框中。此时，选入的变量将作为输入变量，而输出变量则显示为“？”。用户需要在右边的输出变量（Output Variable）框中定义新的变量名及变量标签，单击“更改（Change）”按钮，即可完成输出变量的定义。



图 2.43 “重新编码为其他变量（Recode into Different Variables）”对话框

02 同样，通过“重新编码为其他变量（Recode into Different Variables）”对话框下方的“如果（If）”按钮，用户可以设定个案的条件。单击该按钮，SPSS 将弹出“重新编码到其他变量：如果个案（Recode into Different Variables：If Cases）”对话框。该对话框与图 2.41“重新编码成相同变量：If 个案（Recode into Same Variables：If Cases）”对话框完

全相同，此处不再赘述。

03 选择需要重新分组（编码）的变量后，单击“输入变量→输出变量（Input Variable →Output Variable）”列表框下方的“旧值和新值（Old and New Values）”按钮，将打开“重新编码到其他变量：旧值和新值（Recode into Different Variables：Old and New Values）”对话框，如图 2.44 所示。

“重新编码到其他变量：旧值和新值（Recode into Different Variables：Old and New Values）”对话框左侧的“旧值（Old Value）”栏与“重新编码成相同变量：旧值和新值（Recode into Different Variables：Old and New Values）”对话框完全相同。右上方的“新值（New Value）”栏中除了可以填入新的码值、设定系统缺失值外，还添加了“复制旧值”这一选项。另外，此处 SPSS 还提供了数值型数据与字符型数据相互转换的功能，选中“输出变量为字符串”选项，就可以将数值型数据转换为字符型数据；选中“将数值字符串移动为数值”选项，可以将字符型数据转换为数值型数据。

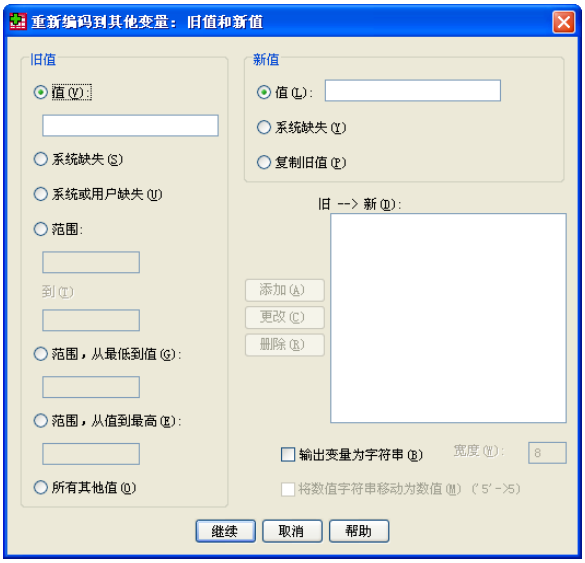


图 2.44　“重新编码到其他变量：旧值和新值（Recode into Different Variables：Old and New Values）”对话框

04 单击“继续”按钮，即可返回“重新编码到其他变量（Recode into Different Variables）”对话框中。在单击“重新编码为其他变量（Recode into Different Variables）”主对话框中单击“确定”按钮，即可执行命令，在数据编辑窗口中得到所需要的新的分组赋值变量。

2.3.3　数据自动分组（编码）

除了通过手动对数据进行重新分组（编码）外，SPSS 还提供了自动对数据进行重新分

组（编码）的功能。但两者的区别在于，前者可以根据用户的需要指定特殊的码值，而后者则是 SPSS 系统自动设定的码（一般为正整数）。数据自动分组（编码）的操作步骤如下。

01 单击“转换（Transform）”菜单下的“自动重新编码（Automatic Recode）”命令，SPSS 将弹出“自动重新编码（Automatic Recode）”对话框，如图 2.45 所示。

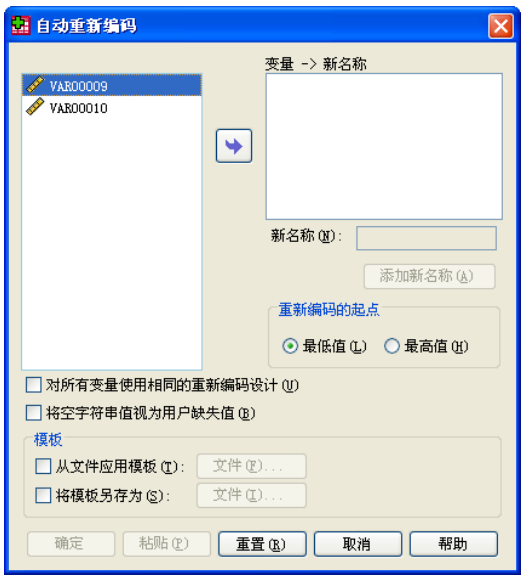


图 2.45 “自动重新编码（Automatic Record）”对话框

02 在“自动重新编码（Automatic Recode）”对话框中，用户可以从对话框左侧的源变量列表中，指定需要进行自动分组的变量，单击 按钮加入“变量→新名称（Variable →New Name）”列表框中。此时，标志变量名显示为“？ ”。用户可以在“新名称（New Name）”右边的空白小窗口中输入新的标志变量名，然后单击“添加新名称（Add New Name）”，即可完成标志变量的定义。

03 在“重新编码的起点（Recode starting from）”栏中，用户可以选择编码是从最低值开始、还是从最高值开始。

04 SPSS 的自动编码还提供了选择模板的功能，用户通过单击“模板（Template）”下的“文件（File）”按钮可以引用自动编码的模板，也可以将现有的设定完毕的自动编码保存为模板。

05 自动编码设定完成后，单击“确定”按钮，即可执行命令，在数据编辑窗口中将得到所需要的标志变量以及原变量响应的分组情况。

2.3.4 产生计数变量

在进行数理统计时，常常需要针对满足一定条件的个案数据进行计数，计算若干个变量中有几个变量的值落在指定的区间内，并将计数结果保存在一个指定的新变量中。例如，

计算某公司每个员工 1~12 月份中满足业绩在 50000 元以上的月份个数。通过计数变量，可以对每个个案满足条件的变量个数有更加直观的认识。产生计数变量的操作步骤如下。

01 单击“转换（Transform）”菜单下的“对个案内的值计数（Count Values within Cases）”命令，SPSS 将弹出“计算个案内值的出现次数（Count Occurrences of Values within Cases）”对话框，如图 2.46 所示。

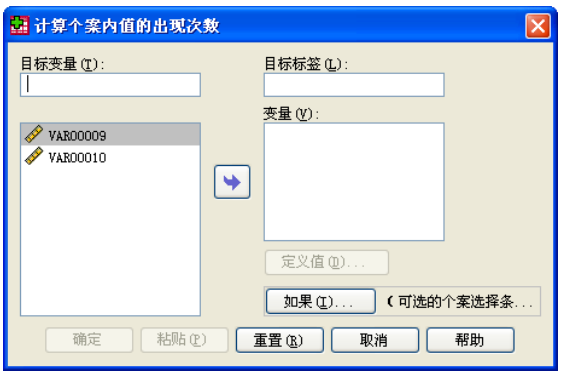


图 2.46 “计算个案内值的出现次数（Count Occurrences of Values within Cases）”对话框

02 用户需要首先在“目标变量（Target Variable）”中指定一个变量作为存放计数结果的变量，并定义变量标签。目标变量既可以是已经存在的变量，也可以是新变量。当目标变量是已经存在的变量时，将会把变量的原数值覆盖为计数结果。接着，用户可以从对话框左侧的源变量列表中，指定需要进行统计的变量，单击 按钮加入“变量（Numeric Variables）”列表框中。单击“定义值（Define Values）”按钮，将打开“统计个案内的值：要统计的值（Count Values within Cases：Values to Count）”对话框，如图 2.47 所示。



图 2.47 “统计个案内的值：要统计的值（Count Values within Cases：Values to Count）”对话框

在“统计个案内的值：要统计的值（Count Values within Cases：Values to Count）”对话框中，可以确定需要计数的数值。

“值（Value）”栏的各选项依次如下。

- 值（Value）：输入某个值为计数对象，等于该值的变量将被计数。
- 系统缺失（System-missing）：以系统的缺失值作为计数对象。
- 系统或用户缺失（System- or user-missing）：以系统或用户所指定的缺失值作为计数对象。
- 范围（Range）：用户可以指定数值的计数区域，在该区间内的数值将作为计数对象。在下面两个小方框内可以指定区间的下限和上限；

范围，从最低到值（LOWEST through value）：在小方框内指定区间的上限；

范围，从值到最高（value through HIGHEST）：在小方框内指定区间的下限。

确定了计数数值（或区间）后，单击“添加（Add）”按钮，可以使计数数值（或区间）进入“要统计的值（Values to count）”列表框中。全部计数数值（或区间）设定完毕后，单击“继续”按钮，即可返回“个案内的数值计数（Count Occurrences of Values within Cases）”主对话框中。

03 如果用户需要设定计数条件，可以通过“计算个案内值的出现次数（Count Occurrences of Values within Cases）”对话框下方的“如果”按钮。单击该按钮，SPSS 将弹出“频数发生：If 个案（Count Occurrences：If Cases）”对话框。该对话框与图 2.41“重新编码成相同变量：If 个案（Recode into Same Variables：If Cases）”对话框完全相同，此处不再赘述。

04 计数变量的条件设定完毕后，单击“确定”按钮，即可执行命令，在数据编辑窗口中将得到新添加的计数变量。

2.3.5 数据秩（序）的确定

在许多统计分析中，都会用到非参数统计的方法，即不考虑总体数据分布类型是否已知，不比较总体数据参数，而只比较总体分布中个案的位置是否相同的统计方法。在非参数统计中，秩和检验是比较常用的。所谓“秩”，就是等级，即按照数据大小排定的次序号。数理统计中，常常需要对个案进行排秩。SPSS 软件的“个案排秩（Rank Cases）”命令就提供了这一功能。

01 单击“转换（Transform）”菜单下的“个案排秩（Rank Cases）”命令，SPSS 将弹出“个案排秩（Rank Cases）”对话框，如图 2.48 所示。

在“个案排秩（Rank Cases）”对话框中，用户可以从对话框左侧的源变量列表中，指定需要进行排秩的变量，单击 按钮加入“变量（Variables）”列表框中。同时，如果需要将数据分组后进行排秩，用户还需选择分组变量，单击 按钮加入“排序标准”列表框中。例如，用户可以以男女为分组变量对某公司员工工资进行排秩。需要注意的是，由于排秩针对的是连续型数值变量；因此，字符型变量将不显示在源变量列表中。



图 2.48　“个案排秩（Rank Cases）”对话框

对话框左下方的“秩 1 指定给（Assign Rank 1 to）”一栏用于选择将秩次“1”赋予最小值还是最大值。

对话框右下方的“显示摘要表（Display summary tables）”选项用于确定是否形成说明的结果文件，该选项默认为“是”。如果取消选择，SPSS 将不再在结果文件中显示摘要表说明，而只在原数据编辑窗口中显示新的秩变量。

02 单击“秩的类型（Rank Types）”按钮，SPSS 将弹出“个案排秩：类型（Rank Cases：Types）”对话框，如图 2.49 所示。



图 2.49　“个案排秩：类型（Rank Cases: Types）”对话框

在“个案排秩：类型（Rank Cases：Types）”对话框中，用户可以定义排秩方式。主要的排秩方式如下。

秩（Rank）：普通排序，即选择后数据编辑窗口中的新变量的值即为对应个案的秩。此选项为系统默认选项。

%分数秩（Fractional rank as %）：累计百分数排序，以百分比的形式呈现，秩值等于普通的秩除以有效个案数再乘以 100。

Savage 得分（Savage score）：以指数分布规律为基础计算出的秩。

个案权重总和（Sum of case weights）：对不同分组的个案总数进行加权。

分数秩（Fractional rank）：分数形式的秩，等于普通的秩除以不同分组的个案总数的加权数。

Ntiles：用户需要在后面的小窗口中给定一个大于 1 的整数，系统按照百分比分组的秩和此数的范围确定秩。

比例估计（Proportion estimates）：用于评估特定秩参数的累计分布比例。用户选择比例估计（Proportion estimates）后，“比例估计公式（Proportion Estimation Formula）”栏将被激活，显示为可选状态，用户可以选择 SPSS 提供的四种比例估计公式中的一种。

Blom：系统默认选项，计算公式为 $\frac{r-3/8}{n+1/4}$。（2.1）

Tukey：计算公式为 $\frac{r-1/3}{n+1/3}$。（2.2）

Rankit：计算公式为 $\frac{r-1/2}{n}$。（2.3）

Van der Waerden：计算公式为 $\frac{r}{n+1}$。（2.4）

以上四个公式中，n 表示个案的数目，r 是从 1 到 n 的秩次。

正态得分（Normal scores）：以与累计分布比例相对应的 Z 分数作为秩。

排秩方式选择完毕后，单击“继续”按钮，即可返回“个案排秩（Rank Cases）”主对话框中。

03 单击“结（Ties）”按钮，SPSS 将弹出“个案排秩：结（Rank Cases：Ties）”对话框，如图 2.50 所示。

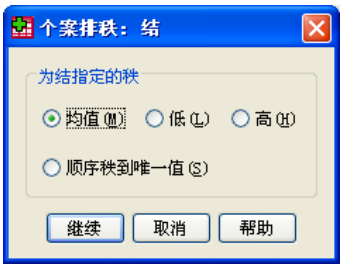


图 2.50 “个案排秩：结（Rank Cases: Ties）”对话框

结（Ties）是指当两个或两个以上的数据相等时秩的处理方式。由于秩与数据排序是一一对应的，当数据相等时，它们对应的秩也相等。在“个案排秩：结点（Rank Cases: Ties）”对话框中，SPSS 提供了下列四种处理相同秩的方式：

- 均值（Mean）：以相等个案秩的均值作为相等个案的秩，例如，3 个相等数据分别排在 60~62 位时，它们对应的秩都为 62；
- 低（Low）：以相等个案秩的最小值作为相等个案的秩，例如，3 个相等数据分别排在 60~62 位时，它们对应的秩都为 60；
- 高（High）：以相等个案秩的最大值作为相等个案的秩，例如，3 个相等数据分别排在 60~62 位时，它们对应的秩都为 62；
- 顺序秩到唯一值（Sequential ranks to unique values）：将数据当作一个连续的记录处理，即 3 个相等数据分别排在 60~62 位时，它们对应的秩按照个案顺序分别为 60、61、62。

等秩处理方式选择完毕后，单击“继续”按钮，即可返回“个案排秩（Rank Cases）”主对话框中。

04 在“个案排秩（Rank Cases）”对话框中，单击“确定”按钮，即可完成全部操作。结果输出窗口中显示的结果说明文件将对新变量进行一定的说明。同时，在数据编辑窗口中，将生成新的变量，即排序结果。

2.3.6　替换缺失值

在进行数理统计分析时，经常需要将带有缺失值的数据剔除后再进行分析。但是有的时候，尤其是在进行时间序列分析时，却不能将缺失值剔除。此时，就需要科学地对缺失值进行处理。SPSS 的“替换缺失值（Replace Missing Values）”提供了替换缺失值的功能，通过替换缺失值并生成新变量达到处理缺失值的目的。

替换缺失值的具体操作步骤如下。

01 单击“转换（Transform）”菜单下的“替换缺失值（Replace Missing Values）”命令，SPSS 将弹出“替换缺失值（Replace Missing Values）”对话框，如图 2.51 所示。



图 2.51　“替换缺失值（Replace Missing Values）”对话框

在“替换缺失值（Replace Missing Values）”对话框中，用户可以从对话框左侧的源变量列表中，指定需要进行缺失值替换的变量，单击 按钮加入“新变量（New Variable(s)）”列表框中。原变量名后将出现为“_1”，以示与原变量的区别，并将其作为自动产生的已替代缺失值的新变量。用户也可以在“名称（Name）”中进行修改，自行定义新变量名。

02 在“方法（Method）”的下拉菜单中，用户可以选择缺失值的替代方式。SPSS 提供了以下五种缺失值的替代方式：

- 序列均值：即以该变量所有非缺失值的平均值替换缺失值。系统默认选择该项；
- 临近点的均值：即以缺失值相邻点的非缺失值的平均数替换缺失值；
- 临近点的中位数：即以缺失值相邻点的非缺失值的中位数替换缺失值；
- 线性插值法：即以缺失值相邻点的非缺失值构建一个数据范围，将两个值之间的差值进行平均：建立数据的等级来作为内插的替换值；
- 点处的线性趋势：即采用最小二乘法对全部数据进行拟合，用拟合得到的数据值替换缺失值。

当用户选择“附近点的均值”和“临近点的中位数”时，“附（邻）近点的跨度（Span of nearby points）”一栏将被激活，用户可以定义相邻点的个数，或者选择全部的变量数据作为相邻点。默认的相邻点个数是 2。

03 在“替换缺失值（Replace Missing Values）”对话框中，单击“确定”按钮，即可完成全部操作。数据编辑窗口中，将生成新的变量，即缺失值已经被替换、非缺失值保持不变。

2.4 数据文件的保存或导出

2.4.1 保存数据文件

如果数据输入编辑完毕，就需要将数据进行保存，方便下次继续使用数据或对数据进行编辑修改。由于 SPSS 能够与多种数据文件进行连接，数据文件也可以保存为不同的类型。但如果今后仍需使用 SPSS 对数据进行统计分析，建议用户将数据保存为 SPSS 默认的.sav 数据格式。

如果该数据从来没有被保存过，单击“文件（File）”菜单下的“保存（Save）”命令，将弹出“将数据保存为（Save Data As）”对话框，如图 2.52 所示。

单击“保存类型（Save as type）”列表框，可以看到 SPSS 所支持的各种数据类型，SPSS 可以保存的数据类型包括：SPSS（.sav、.sys、.syd、.por）、Excel（.xls）、Lotus（.w）、Text（.txt）等。选择数据文件对应的数据类型，并在文件名框内输入新文件名，单击“保存（Save）”按钮，即完成数据保存的操作。返回数据编辑窗口，可以看到标题栏左上角的文件名将由 Untitled 变为了新保存的文件名。

如果用户之前已对数据进行过保存，单击“文件（File）”菜单下的“保存（Save）”

命令，可保存最新进行的数据编辑操作。



图 2.52　“将数据保存为（Save Data As）”对话框

如果用户对数据进行保存后，仍想将新进行的编辑操作后的数据保存至另一文件名下，可以选择“文件（File）”菜单下的“另存为（Save As）”命令，SPSS 也将弹出“将数据保存为（Save Data As）”对话框，操作过程与保存文件一致。

2.4.2　导出数据文件

除了可以直接将数据保存为许多类型的数据文件外，SPSS 还提供了另一个适用范围更广、但使用也更为专业的数据保存方式——导出到数据库。与打开数据库查询一样，SPSS 使用了 ODBC（Open Database Connectivity）数据接口，该接口能被大多数数据库软件和办公软件（如 Microsoft Office）支持，通过 ODBC 应用程序可以直接访问以结构化查询语言（SQL）作为数据访问标准的数据库管理系统。由于 SPSS 17.0 for Windows 中文版可以直接保存除 Excel 2007 以外的所有 Excel 数据文件，因此，导出到数据库对于保存 Excel 文件的用处不是很大。

单击“文件（File）”菜单，选择“导出到数据库（Export to Database）”命令，系统会弹出“导出到数据库向导（Export to Database Wizard）”对话框，如图 2.53 所示。窗口中将列出所用机器上已安装的所有数据库驱动程序，选中所需的数据库类型，根据向导的提示即可将数据导出到指定的数据库。

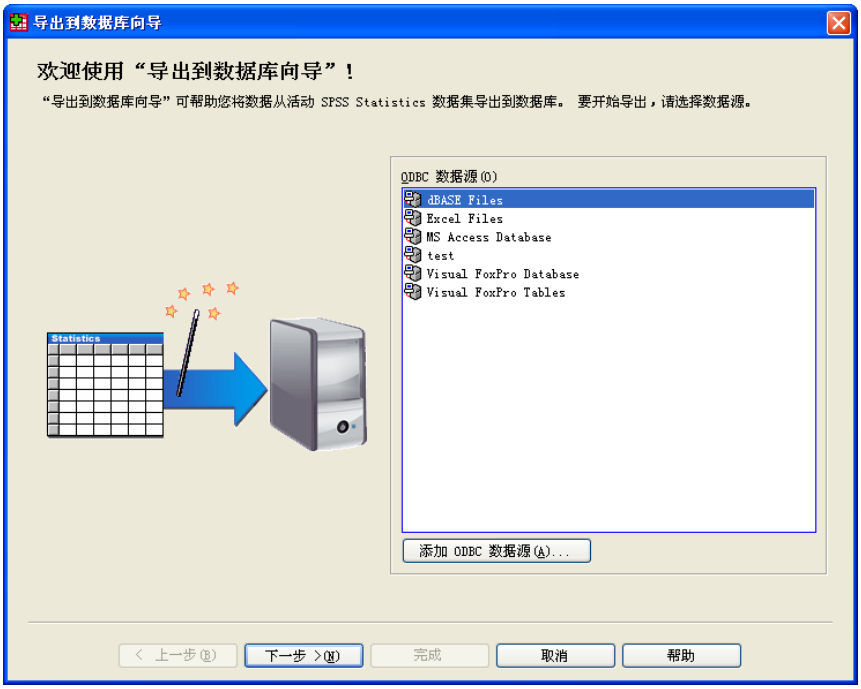


图 2.53 “导出到数据库向导（Export to Database Wizard）”对话框

2.5 本章小结

本章主要介绍了数据编辑管理的方法和过程，包括建立数据文件、编辑加工数据和保存数据文件。本章的重点在于数据的编辑加工部分。通过对本章的学习，读者可以自行建立数据文件，并对数据进行整理、转换，也为正式进入数理统计分析学习铺平道路。

第3章 统计图表绘制

由于单纯的输出数字过于死板，也很难形象地说明数字所代表的意义。因此，有时我们需要绘制统计图表，把资料所反映的变化趋势、数量多少、分布状态和相互关系等形象直观地表现出来，以便于读者的阅读、比较和分析。

统计表图正是用点的位置、线段的升降、直条的长短或面积的大小等来表达资料的内容。SPSS 17.0 的“图表（Graphs）”菜单提供了绘制图表的功能。

SPSS 17.0 的“图表（Graphs）”菜单主要包括三个子菜单。“图表建立（Chart Builder）”相当于图表向导，它对 SPSS 的绘图功能做了粗略地介绍，初学者可以大致了解 SPSS 绘图的能力；“交互图表（Interactive）”主要涵盖了 SPSS 各种复杂的交互性图表；“旧对话框（Legacy Dialogs）”主要包括多种 SPSS 传统的常用的统计报表，这也将是本章重点阐述的内容。

本章学习要点：

- 条形图；
- 线图；
- 面积图
- 饼图；
- 箱图；
- 直方图。

3.1 条形图

条形图（Bar Charts），又称带形图或柱形图，它是利用相同宽度的条形的长短或高低来表现统计数据大小或变动的统计图。

绘制条形图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“条形图

（Bar Charts）”，SPSS 将弹出“条形图（Bar Charts）”导航对话框，如图 3.1 所示。

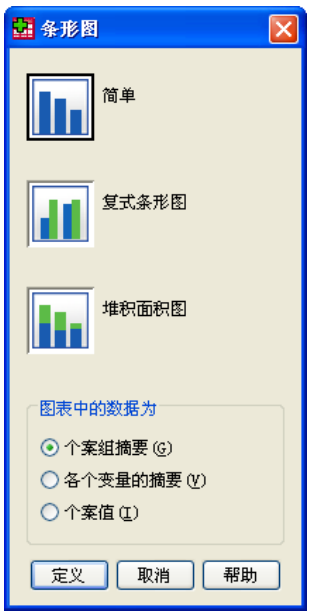


图 3.1 “条形图（Bar Charts）”导航对话框

在“条形图（Bar Charts）”导航对话框中，用户可以选择条形图的类型，并定义条形图中数据的表达方式。

SPSS 将条形图大致分为三种类型：

- 简单（Simple）：单式条形图，各个条形相互独立；
- 复式条形图（Clustered）：分组条形图，每组的相邻两个或多个条形左右连结，以方便进行组间比较；
- 堆积面积图（Stacked）：分段条形图，同一组的条形图上下连结，以方便进行组与组之间的互相比较。

在图 3.1 下部的“图表中的数据为（Data in Chart are）”栏，用户可以选择如下条形图中数据的表达类型：

- 个案组摘要（Summaries for groups of cases）：用分类值作图，条形图中每一条代表个案的一个分类；
- 各个变量的摘要（Summaries of separate variables）：用变量值作图，条形图中每一条代表一个变量；
- 个案值（Values of individual cases）：用单元值作图，条形图中每一条代表一个观察值。

通过以上三个条形图类型和三个数据表达方式的不同搭配，SPSS 可以生成 9 种不同的条形图。本书以用户选择“简单（Simple）”条形图和“个案组摘要（Summaries for groups of cases）”为例，阐述条形图的绘制步骤。

02 单击“定义（Define）”按钮，进入正式的条形定义对话框“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for groups of cases）”对话框，如图 3.2 所示。根据用户所选的条形图类型和数据表达方式的不同，出现的对话框名称也不同，但对话框的主体内容大致相同。

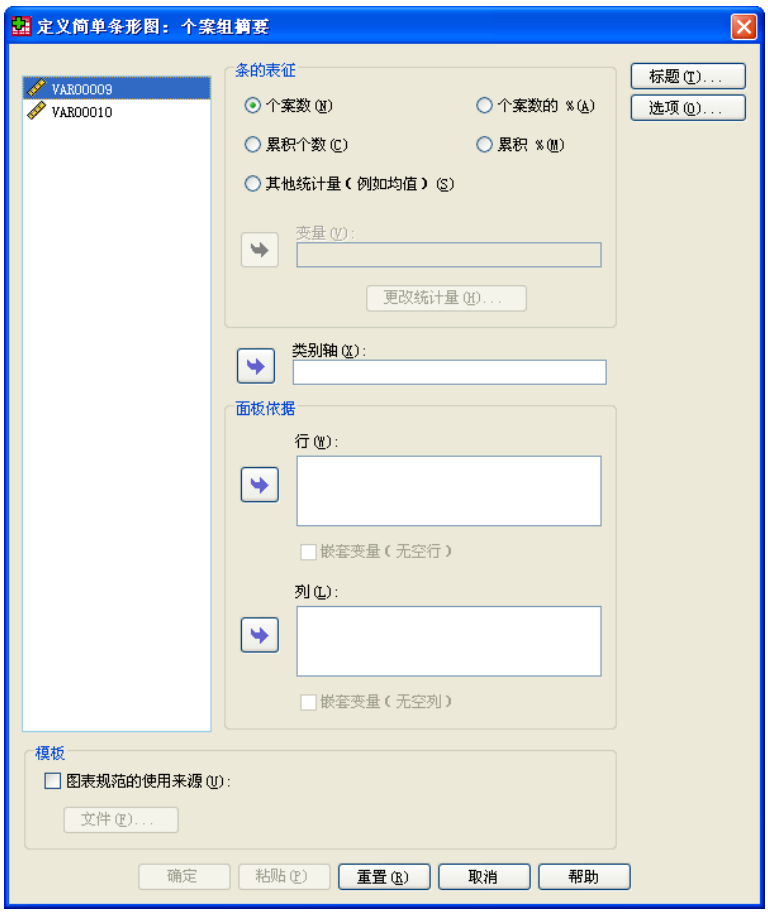


图 3.2 “定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框

在“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框中，用户可以选择条形图绘制的相关细节。

“条的表征（Bar Present）”栏，用户可以选择以下条形图中条所代表的统计量：

- 个案数（N of cases）：按照分组变量分组后各组的个案个数；
- 个案数的%（% of cases）：按照分组变量分组后各组的个案个数占总个案个数的百分比；
- 累积个数：个案数的累计数目；
- 累积%：个案数的累计百分比；

- 其他统计量（Other statistics）：用户可以自行定义条形图中的统计量。选中该选项，下面的“变量（Variable）”列表框将被激活，用户需要通过单击从左边源变量中选择一个分析变量进入“变量（Variable）”列表框中，然后单击“更改统计量（Change Statistics）”按钮，SPSS 将弹出“统计量（Statistic）”对话框，如图 3.3 所示。

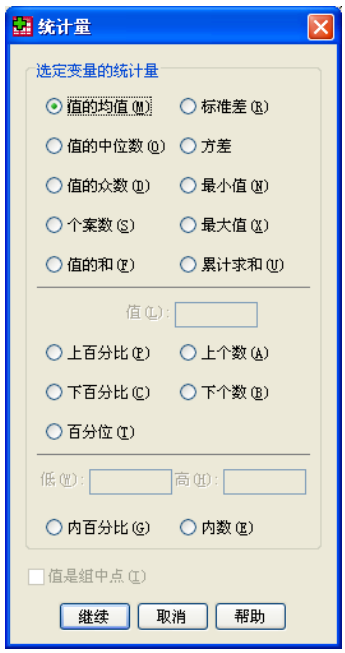


图 3.3 “统计量（Statistic）”对话框

在“统计量（Statistic）”对话框中，用户可以选择需要计算的描述统计量作为条形图中的统计量。对话框中的描述统计量包括：

- 值的均值（Mean of values）：按照分组变量分组后，以分析变量的均值作为条形图中的统计量；
- 值的中位数（Median of values）：按照分组变量分组后，以分析变量的中位数作为条形图中的统计量；
- 值的众数（Mode of values）：按照分组变量分组后，以分析变量的众数作为条形图中的统计量；
- 个案数（Number of cases）：按照分组变量分组后，以分析变量的样本个数作为条形图中的统计量；
- 值的和（Sum of values）：按照分组变量分组后，以分析变量数据的总和作为条形图中的统计量；
- 标准差（Standard deviation）：按照分组变量分组后，以分析变量的标准差作为条

形图中的统计量；

- 方差（Variance）：按照分组变量分组后，以分析变量的方差作为条形图中的统计量；
- 最小值（Minimum value）：按照分组变量分组后，以分析变量的最小值作为条形图中的统计量；
- 最大值（Maximum value）：按照分组变量分组后，以分析变量的最大值作为条形图中的统计量；
- 累计求和（Calculative sum）：按照分组变量分组后，以分析变量的累计总和作为条形图中的统计量。

在该对话框的中部，用户可以在“值（Value）”窗口中输入数值或者百分数，然后选择相关选项，SPSS 将对数值或百分数上侧或下侧的个案按照大小进行筛选。

在该对话框的下部，用户可以在“低（Low）”和“高（High）”窗口中输入数值或者百分数，然后选择对应选项，SPSS 将按照大小对个案在限定范围内进行筛选。

对话框底部的“值是组中点（Values are grouped midpoints）”复选项，只有在用户选择值的中位数（Median of values）或者“百分位（Percentile）”时，才被激活。选中该选项，则表明数据分布为频数分布表的格式，输出的条形图中的统计量为分组的中值。

选择条形图中的统计量后，单击“继续”按钮，即可返回到“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”主对话框中。

接下来，需要指定作图变量。在“类别轴（Category Axis）”列表框中，用户需要通过单击从左边源变量中选择一个变量作为分类变量（也称分组变量）。条形图中每个条形的长度分布代表对应各组的统计量的值。

模板（Template）：用户可以选中该选项，并单击“文件（File）”按钮，选择模板文件，作为散点图的格式模板。

03 单击“标题（Titles）”按钮，打开“标题（Titles）”对话框，如图 3.4 所示。

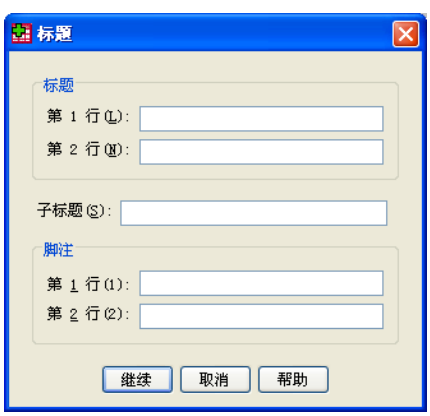


图 3.4 “标题（Titles）”对话框

在“标题（Titles）”对话框中，用户可以定义散点图的标题、子标题和脚注。

标题（Title）和子标题（Subtitle）：通过输入，用户可以定义两行标题和一行子标题。标题和子标题将在散点图上方显示。用户也可以在结果输出窗口中定义标题。

脚注（Footnote）：通过输入，用户可以定义两行脚注，脚注将在散点图下方显示。

输入标题后，单击“继续”按钮，即可返回“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”主对话框。

04 单击“选项（Options）”按钮，打开“选项（Options）”对话框，如图 3.5 所示。

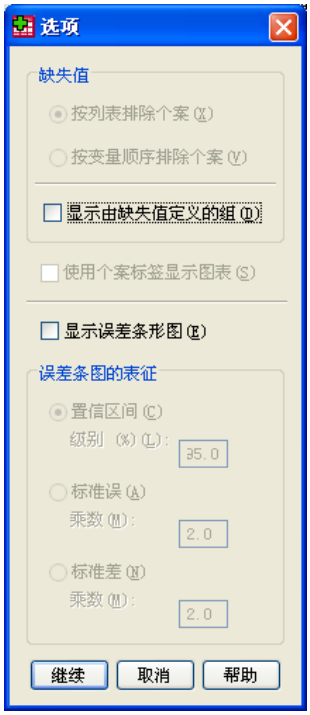


图 3.5 “选项（Options）”对话框

在“选项（Options）”对话框中，用户可以指定缺失值的处理方式和误差条的设定形式。

缺失值（Missing Values）栏：用户可以定义分析中对缺失值的处理方式，包括按列表排除个案（Exclude cases listwise）和按变量顺序排除个案（Exclude cases variable by variable）两种方式。为了尽可能充分利用数据，SPSS 默认选择前者。

误差条图的表征（Error Bars Represent）栏：用户可以选择设定置信区间（Confidence intervals）、标准误（Standard Error）或者标准差（Standard deviation）条件，并显示误差条图。

指定缺失值的处理方式和误差条的设定形式后，单击“继续”按钮，即可返回“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”主对话框。

05 在“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到简单条形图。

3.2　线图

线图（Line Charts），又称曲线图，它是利用点的高低来表明数据升降情况的一种统计图。通过将不同阶段的数据点连接起来，可以更清晰地表明数据的变化趋势。线图主要用于时间序列分析、数据分配情况比较和两变量依存关系的分析等。

绘制线图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“线图（Line Charts）”，SPSS 将弹出“线图（Line Charts）”导航对话框，如图 3.6 所示。



图 3.6　“线图（Line Charts）”导航对话框

在“线图（Line Charts）”导航对话框中，用户可以选择线图的类型，并定义线图中数据的表达方式。

SPSS 将线图大致分为三种类型：

- 简单（Simple）：单线图，一个图形中只有一条水平走向的折线；
- 多线线图（Multiple）：多线图，一个图形中有多条水平走向的折线；
- 垂直线图（Drop-line）：垂线图，一个图形中有多组水平走向的数据，但在水平方向上不予以连接，而只是在垂直方向上将同一时间点的数据予以连接。

“图表中的数据为（Data in Chart are）”栏：用户可以选择以下线图中数据的表达类型：

- 个案组摘要（Summaries for groups of cases）：用分类值作图，线图中每一条代表个案的一个分类；

- 各个变量的摘要（Summaries of separate variables）：用变量值作图，线图中每一条代表一个变量；
- 个案值（Values of individual cases）：用单元值作图，线图中每一条代表一个观察值。

通过以上三个线图类型和三个数据的表达类型的不同搭配，SPSS 可以生成 9 种不同的线图。本书以用户选择“多线（Multiple）”线图和“个案组摘要（Summaries for groups of cases）”为例，阐述线图的绘制步骤。

02 单击“定义（Define）”按钮，进入正式的线图定义对话框“定义多线线图：个案组摘要（Define Multiple Line：Summaries for groups of cases）”对话框，如图 3.7 所示。根据用户所选的线图类型和数据表达类型的不同，出现的对话框名称也不同。

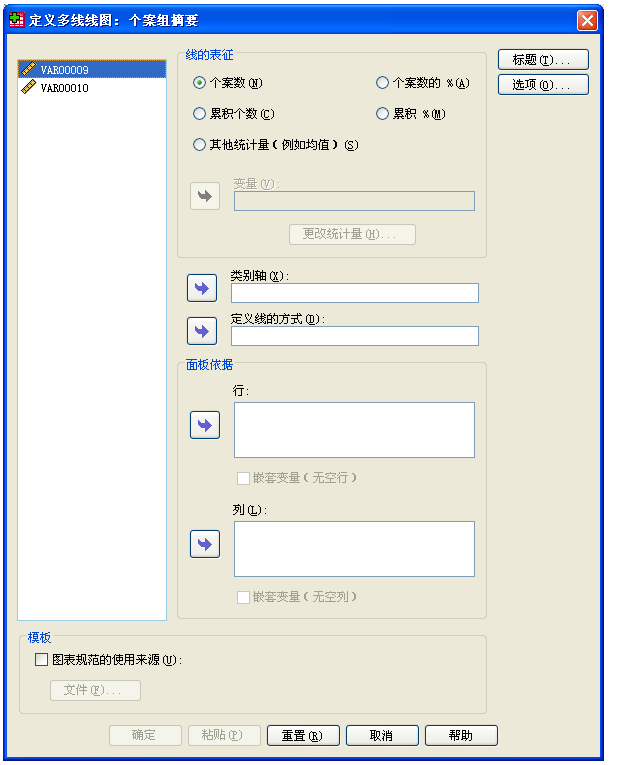


图 3.7 “定义多线线图：个案组摘要（Define Multiple Line：Summaries for groups of cases）”对话框

在“定义多线线图：个案组摘要（Define Multiple Line：Summaries for groups of cases）”对话框中，用户首先需要指定绘图变量，即通过单击从左边源变量中选择多个需要绘制折线图的变量进入右边的“变量”列表框中。绘图变量的数值将在线图的纵轴上表示。

同时，用户需要指定分类变量。在“线的表征（Lines Represent）”栏中，用户可以选择以“个案数（Case number）”，即个案的编号作为分类变量。用户也可以选择“变量（Variable）”选项，然后单击选择一个变量作为分类变量。例如，在时间序列分析中，用户就可以将时间变量作为分类变量。分类变量的数值将在线图的横轴上表示。

“定义多线线图：个案组摘要（Define Multiple Line：Summaries for groups of cases）”对话框的其他部分以及单击“标题（Title）”按钮所弹出的“标题（Title）”对话框都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，此处不再赘述。

03 在“定义多线线图：个案组摘要（Define Multiple Line：Summaries for groups of cases）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到多线图。

3.3　面积图

面积图（Area Charts），又称区域图，是用面积来表现连续性的频数分布的统计图。面积越大，频数越多，反之亦然。制作面积图的原始数据大多为频数分布表。

绘制面积图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“面积图（Area Charts）”命令，SPSS 将弹出“面积图（Area Charts）”导航对话框，如图 3.8 所示。



图 3.8　“面积图（Area Charts）”导航对话框

在“面积图（Area Charts）”导航对话框中，用户可以选择面积图的类型，并定义面积图中数据的表达方式。

SPSS 将面积图大致分为以下两种类型：

- 简单（Simple）：简单面积图，图形中只有一个区域表示频数大小；
- 堆积面积图（Stacked）：图形由多个不同颜色的区域从下到上堆积而成，不同的区域表示不同类型的频数大小。

“面积图（Area Charts）”导航对话框中的“图表中的数据为（Data in Chart are）”栏与条形图、线图的导航对话框该栏的内容相同，这里不再赘述。

通过以上两个面积图类型和三个数据的表达类型的不同搭配，SPSS 可以生成 6 种不同的面积图。本书以用户选择“堆积（Stacked）面积图”和“个案组摘要（Summaries for groups of cases）”为例，阐述面积图的绘制步骤，如果用户需要制作其他类型的面积图，请参看堆积面积图绘制的类似步骤。

02 单击“定义（Define）”按钮，进入“定义堆栈面积：个案组摘要（Define Stacked Area：Summaries for groups of cases）”对话框，如图 3.9 所示。根据用户所选的面积图类型和数据表达方式的不同，出现的对话框名称也不同。

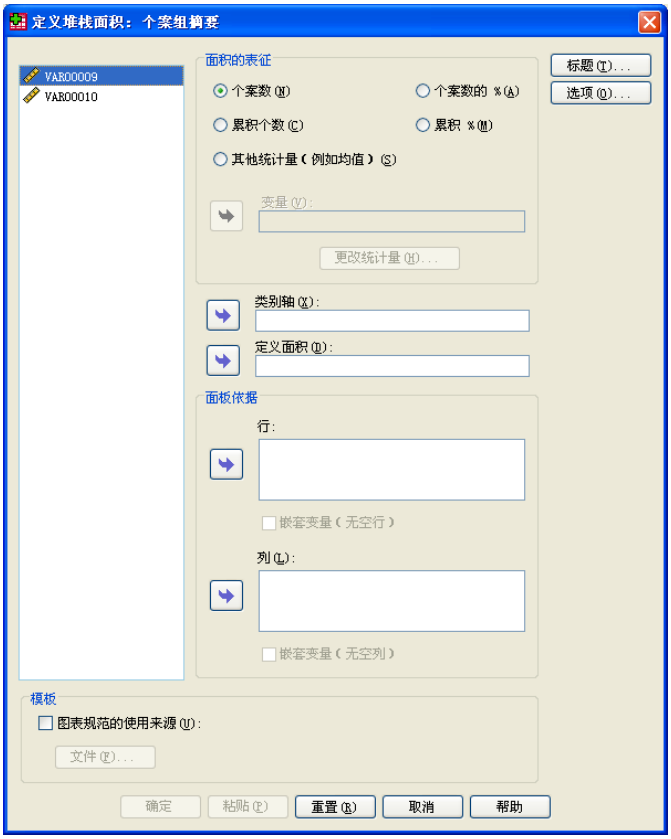


图 3.9 “定义堆栈面积：个案组摘要（Define Stacked Area：Summaries for groups of cases）”对话框

“定义堆栈面积：个案组摘要（Define Stacked Area：Summaries for groups of cases）”对话框与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框几乎完全相同，只是在“类别轴（Category Axis）”下方多了一个“定义面积（Define Areas by）”列表框。在该列表框中，用户需要单击从左边源变量列表中选择一个变量作为面积的识别变量。例如，将不同性别的学生的成绩频数分布表用面积图显示时，需要将男女生分别用两个不同颜色的区域区分开来，此时，就可以将性别作为面积的

识别变量。

该对话框的其他部分以及“标题（Title）”按钮、“选项（Options）”按钮都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，此处不再赘述。

03 在“定义堆栈面积：个案组摘要（Define Stacked Area：Summaries for groups of cases）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到堆积面积图。

3.4　饼图

饼图（Pie Charts）是用扇形的大小来表示数值大小的统计图。饼图简单易懂，并且便于比较，因此，饼图的制图过程使用最为广泛。

绘制饼图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“饼图（Pie Charts）”命令，SPSS 将弹出“饼图（Pie Charts）”导航对话框，如图 3.10 所示。

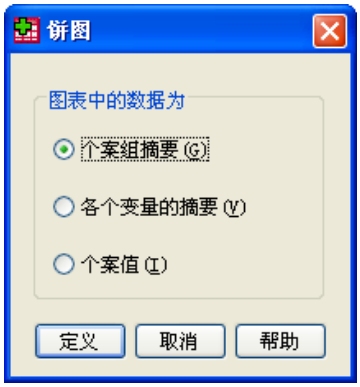


图 3.10　“饼图（Pie Charts）”导航对话框

在“饼图（Area Charts）”导航对话框中，用户可以定义饼图中数据的表达方式。

“饼图（Area Charts）”导航对话框中只有“图表中的数据为（Data in Chart are）”一栏，并且与条形图、线图、面积图的导航对话框该栏的内容相同，此处不再赘述。

根据三个不同的数据表达方式，SPSS 可以生成 3 种不同的面积图。本书以用户选择“个案组摘要（Summaries for groups of cases）”为例，阐述饼图的绘制步骤，如果用户需要制作其他类型的饼图，请举一反三学习类似的步骤。

02 单击“定义（Define）”按钮，进入“定义饼图：个案组摘要（Define Pie：Summaries for groups of cases）”对话框，如图 3.11 所示。根据用户所选的数据表达方式的不同，出现的对话框名称也不同。

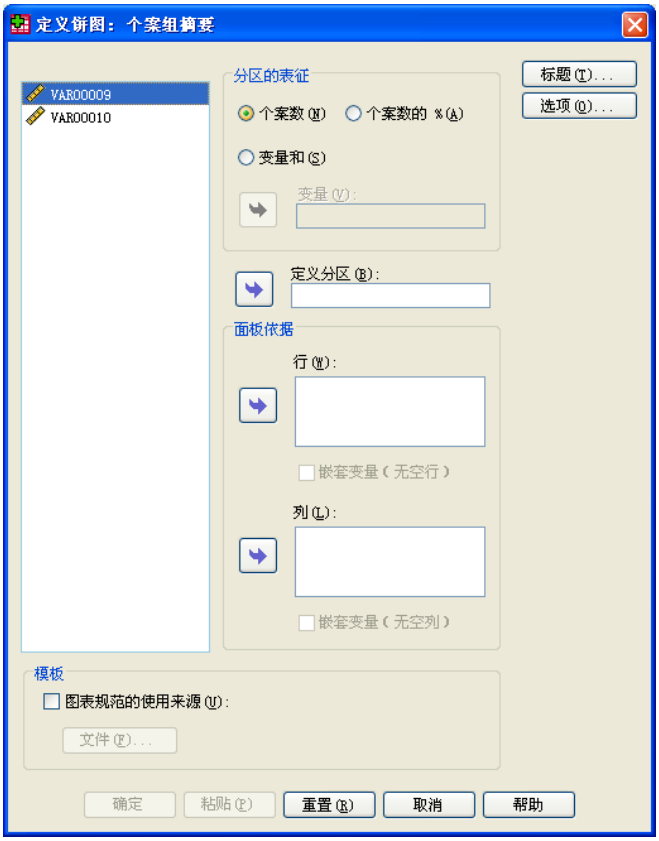


图 3.11 “定义饼图：个案组摘要（Define Pie：Summaries for groups of cases）”对话框

在“定义饼图：个案组摘要（Define Pie：Summaries for groups of cases）”对话框中，用户可以选择饼图绘制的相关细节。

“分区的表征（Bar Present）”栏：用户可以选择饼图中扇形切片所代表的统计量。

- 个案数（N of cases）：按照分组变量分组后各组的个案个数。
- 个案数的%（% of cases）：按照分组变量分组后各组的个案个数占总个案个数的百分比。
- 变量和（Sum of variable）：按照分组变量分组后，用户可以选择另一变量并以该变量的总和作为切片统计量的表达方式。选中该选项后，下面的“变量（Variable）”列表框被激活，用户需要通过单击从左边源变量中选择一个计算总和的变量进入“变量（Variable）”列表框中。

在“定义分区（Define Slices by）”列表框中，用户需要单击从左边源变量列表中指定一个分类变量作为不同扇形的识别变量。

该对话框的其他部分以及“标题（Title）”、“选项（Options）”按钮都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，

此处不再赘述。

03 在“定义饼图：个案组摘要（Define Pie：Summaries for groups of cases）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到饼图。

3.5　高低图

高低图（High-Low Charts）是用多个垂直线段来表示数值区域的统计图，例如一组测定值的范围（最小值-最大值）、95%置信区间（下限-上限）、$\overline{X}$ ±1.96 • SD（低值-均值-高值）等。

绘制高低图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“高低图（High-Low Charts）”命令，SPSS 将弹出“高–低图（High-Low Charts）”导航对话框，如图 3.12 所示。



图 3.12　“高-低图（High-Low Charts）”导航对话框

在“高–低图（High-Low Charts）”导航对话框中，用户可以选择高低图的类型，并定义高低图中数据的表达方式。

SPSS 将高低图大致分为如下五种类型：

- 简单高低关闭（Simple high-low-close）：一个图形中，只有一组高低图，高值与低值之间用垂直线段连接。闭合（close）是指用户指定的一个特殊变量，特殊变量的数值可以在图中用小圆圈的形式标识出来；

- 群集高低关闭（Clustered high-low-close）：一个图形中，有多组高低图，每组高值与低值之间用垂直线段连接，不同组别的高低图用不同颜色的线段表示；
- 简单范围栏（Simple range bar）：一个图形中，只有一组高低图，高值与低值之间用垂直条形连接；
- 群集范围栏（Clustered range bar）：一个图形中，有多组高低图，每组高值与低值之间用垂直条形连接，不同组别的高低图用不同颜色的条形表示；
- 差别面积（Difference）：一个图形中，只有一组高低图，根据分组变量，对应的高值与高值连接，低值与低值连接，两条折线中形成面积区域。

“高-低图（High-Low Charts）”导航对话框中的“图表中的数据为（Data in Chart are）”栏与其他图形的导航对话框该栏的内容相同，此处不再赘述。

通过以上 5 个高低图类型和 3 个数据的表达方式的不同搭配，SPSS 可以生成 15 种不同的面积图。本书以用户选择“简单高低关闭（Simple high-low-close）”和“各个变量的摘要（Summaries of separate variables）”为例，阐述高低图的绘制步骤，如果用户需要制作其他类型的面积图，请参看简单高低图绘制的类似步骤。

02 单击“定义（Define）”按钮，进入正式的对话框“定义高-低-闭合：各个变量的摘要（Define Simple high-low-close：Summaries of separate variables）”对话框，如图 3.13 所示。根据用户所选的高低图类型和数据表达方式的不同，出现的对话框名称也不同。

在“定义高-低-闭合：各个变量的摘要（Define Simple high-low-close：Summaries of separate variables）”对话框中，用户可以选择高低图绘制的相关细节。

用户可以通过单击从左边源变量中选择如下三个变量进入右边对应的变量列表框中：

- “高（High）”：作为高值的变量；
- “低（Low）”：作为低值的变量；
- “闭合（Close）”：用户需要指定一个变量，该变量的数值可以在图中以小圆圈的形式标识出来。

以上三个变量，用户都可以选择指定的统计量作为显示的数值，只要在对应列表框中选中变量，然后单击“更改统计量（Change Statistics）”，SPSS 就会弹出“统计（Statistic）”对话框。默认的统计量是“数据的平均值（Mean of values）”。

在“类别轴（Category Axis）”列表框中，用户同样需要通过单击从左边源变量中选择一个变量作为分类变量（也称分组变量），以保证高值、低值、闭合值的一一对应。

该对话框的其他部分以及“标题（Title）”按钮、“选项（Options）”按钮都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，此处不再赘述。

03 在“定义高-低-闭合图：各个变量的摘要（Define Simple high-low-close：Summaries of separate variables）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到高低图。

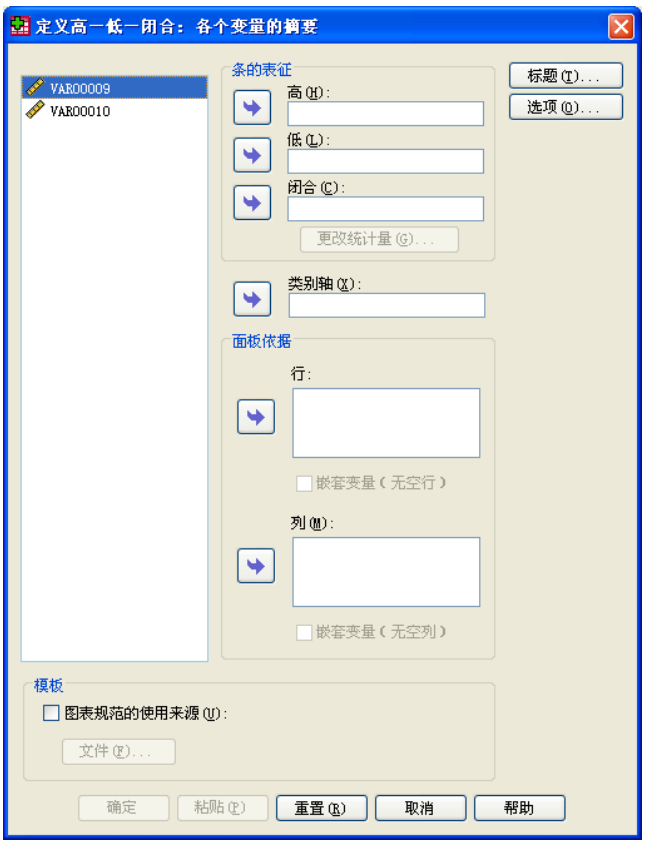


图 3.13　“定义高-低-闭合：各个变量的摘要（Define Simple high-low-close：Summaries of separate variables）”对话框

3.6　箱图

箱图（Boxplot），又称箱锁图，是表示样本数据的中位数、四分位数和极端值的统计图形。在箱图中，最上方和最下方的线段分别表示数据的最大值和最小值，其中箱型的上方和下方的线段分别表示第三四分位数和第一四分位数，箱型中间的粗线段表示数据的中位数。另外，箱图中在最上方和最下方的星号和圆圈分布表示样本数据中的极端值。

绘制箱图的具体操作步骤如下。

01 打开“图表(Graphs)”菜单，选择“旧对话框(Legacy Dialogs)”下的“箱图(Boxplot)”命令，SPSS 将弹出“箱图（Boxplot）”导航对话框，如图 3.14 所示。

在“箱图（Boxplot）”导航对话框中，用户可以选择箱图的类型，并定义箱图中数据的表达方式。

SPSS 将箱图大致分为以下两种类型：

- 简单（Simple）：一个图形中有多个箱型，各个箱型相互独立；
- 复式（Clustered）条形图：一个图形中有多个箱型，多个箱型之间按照分组变量分成若干组，相同组别的箱型集中放置，以方便用户进行比较。



图 3.14 “箱图（Boxplot）”导航对话框

“图表中的数据为（Data in Chart are）”栏：用户可以选择以下箱图中数据的表达类型：

- 个案组摘要（Summaries for groups of cases）：用分类值作图，箱图中每一条代表个案的一个分类；
- 各个变量的摘要（Summaries of separate variables）：用变量值作图，箱图中每一条代表一个变量。

通过以上两个箱图类型和两个数据的表达方式的不同搭配，SPSS 可以生成 4 种不同的箱图。本书以用户选择“简单（Simple）”箱图和“个案组摘要（Summaries for groups of cases）”为例，阐述箱图的绘制步骤，如果用户需要制作其他类型的箱图，请参看简单箱图绘制的类似步骤。

02 单击“定义（Define）”按钮，进入正式的箱型定义对话框“定义简单箱图：个案组摘要（Define Simple Boxplot：Summaries for groups of cases）”对话框，如图 3.15 所示。根据用户所选的箱图类型和数据表达方式的不同，出现的对话框名称也不同，但对话框的主体内容大致相同。

在“定义简单箱图：个案组摘要（Define Simple Boxplot：Summaries for groups of cases）”对话框中，用户可以选择以下箱图绘制的相关细节：

- “变量（Variable）”列表框：用户需要通过单击从左边源变量中选择一个变量作为分析变量，分析变量的数据将作为箱图中各四分位点的依据；
- “类别轴（Category Axis）”列表框：用户需要选择一个变量作为分类变量（也称分组变量），SPSS 将按照分组变量所分的组别决定箱图的个数；

- “标签个案依据（Label Cases by）”列表框：非必选项，用户可以选择一个变量作为标识变量，以保证分析变量与分类变量的一一对应。

该对话框的其他部分以及“选项（Options）”按钮所弹出的“选项（Options）”对话框都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，此处不再赘述。



图 3.15　“定义简单箱图：个案组摘要（Define Simple Boxplot：Summaries for groups of cases）”对话框

03 在“定义简单箱图：个案组摘要（Define Simple Boxplot：Summaries for groups of cases）”主对话框中，单击“确定”按钮，即可在结果输出窗口得到箱图。

3.7　直方图

直方图（Histogram）是用条形的长短来表示连续性频数大小的统计图。其统计意义与面积图（Area Charts）相似，但是区域图能进行多组资料的比较（通过堆积式区域图），而直方图无此项功能。另外，由于直方图也是用条形的长短来表示统计意义的，故其外观上与条形图类似。但由于条形图是分析非连续频数，而直方图是分析连续性频数，因此，条形图的条形之间存在一定间隔，而直方图的条形之间不存在间隔。

绘制直方图的具体操作步骤如下。

01 打开“图表（Graphs）”菜单，选择“旧对话框（Legacy Dialogs）”下的“直方图（Histogram）”命令，SPSS 将弹出“直方图（Histogram）”对话框，如图 3.16 所示。

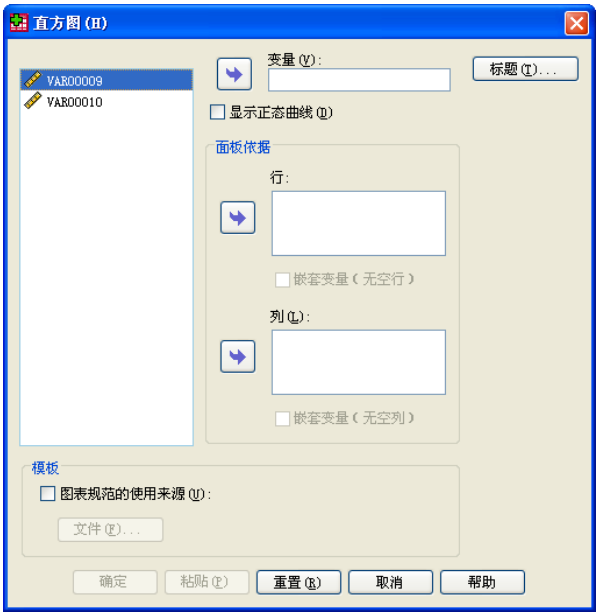


图 3.16 “直方图（Histogram）”对话框

02 在“直方图（Histogram）”对话框中，用户可以选择直方图绘制的相关细节。

用户需要通过单击从左边源变量中选择一个变量作为分析变量进入“变量（Variable）”列表框中。SPSS 自动将分析变量进行频数分析，并根据各个频数分布段绘制直方图。直方图的横轴代表分析变量数据的频数区间，纵轴代表每个区间的频数。

“显示正态曲线（Display normal curve）”复选框：选择该选项，SPSS 将在频数分布图中绘制正态分布曲线，以方便用户将数据与正态分布进行比较，判断样本数据是否符合正态分布。

该对话框的其他部分以及“标题（Title）”按钮都与“定义简单条形图：个案组摘要（Define Simple Bar：Summaries for Groups of Cases）”对话框完全相同，此处不再赘述。

03 在“直方图（Histogram）”对话框中，单击“确定”按钮，即可在结果输出窗口得到直方图。

3.8 本章小结

本章详细介绍了各种统计图表的具体操作步骤。由于图表本身具有易于比较、展示数字关系的特点，因此，在统计分析中，SPSS 的“图表（Graphs）”菜单使用极为广泛，常常与“分析（Analyze）”菜单的统计功能搭配使用。例如，进行频数分析时绘制直方图，进行探索性分析时绘制箱型图，进行时间序列分析时绘制折线图等，读者可在以后的实践应用中具体体会。

第4章 基础统计描述

当样本数据采集完全之后，常用一些数学统计量来描述原始数据的集中程度、离散状况和分布情况，通过这些统计量，我们可以对数据的总体特征进行归纳。

SPSS 基本统计分析是进行其他统计分析的基础和前提。通过基本统计分析方法，可以对要分析数据的总体特征有比较准确的把握，从而可以选择其他更为深入的统计分析方法。

本章学习要点：

- 数据描述；
- 频数分析；
- 探索分析；
- 交叉列联表分析；
- 比率分析；
- P-P 图、Q-Q 图。

4.1 数理统计量概述

SPSS 在描述性统计分析中，提供了多个统计量来描述数据特征，这些统计量包括均值、中位数、众数、方差、标准差、四分位数、十分位数、百分位数、峰度系数、偏度系数等。在进行描述性统计分析之前，首先要对这些统计量在统计学上的定义及其计算公式有所了解。

4.1.1 均值（Mean）和均值标准误差（S.E. mean）

均值（平均数、平均值）表示的是某个变量所有取值的集中趋势或平均水平。例如，某班学生数学考试的平均成绩、公司员工的平均收入、某年级学生的平均身高、某高校高招录取平均分等。

平均数有总体平均数和样本平均数之分。

总体平均数。若一组数据X_1，X_2，…，X_N代表一个大小为N的有限总体，则其总体平均数为：

$$\mu=\frac{\sum_{i=1}^{N}X_i}{N} \tag{4.1}$$

样本平均数。若一组数据x_1，x_2，…，x_n代表一个大小为n的有限样本，则其样本平均数为：

$$\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{4.2}$$

样本数据是从总体数据中抽取出来的，但在不同次抽样中，得到的样本是不同的。虽然在一定程度上，样本数据可以反映总体数据的特征，但由于抽样等原因，样本数据是总体数据的随机变量。同样地，虽然样本均值可以反映总体数据的特征，但在不同次抽样中，所得的样本均值是不同的，并且它们与总体均值间存在不同的差异。

均值标准误差（Standard Error of Mean，S.E. mean，简称标准误）就是描述这些样本均值与总体均值之间平均差异程度的统计量。

4.1.2 中位数（Median）

中位数是将总体数据的各个数值按大小顺序排列，居于中间位置的变量，用 Median 表示。中位数将所有的数据等分成两半，中位数两端的数据个数相同，因此它也被称为二分位数。中位数的确定，仅仅取决于它在数列中的位置，不受极端值的影响；因此，可以用它表示总体的一般水平。同时，中位数比算术平均数具有更好的稳定性。

一个大小为 N 的数列，要求其中位数，首先应把该数列按大小顺序排列，如果 N 为奇数，那么该数列的中位数就是$\frac{N+1}{2}$位置上的数；如果 N 为偶数，中位数则是该数列中第$\frac{N}{2}$与第$\frac{N}{2}+1$位置上的两个数值的平均数。

4.1.3 众数（Mode）

众数是指总体数据中出现次数最多的变量，用 Mode 表示。它同样不受数据极端值的影响，从而在一定程度上提高了平均水平的代表性。例如，制衣厂可以根据消费者的所需服装尺码的众数来安排生产。此外，如果众数的值出现的频数或频率较大，那么说明此众数的代表性就越高，数列的集中趋势也就越显著。

确定众数没有明确的公式，一般只能用手工统计，故较为烦琐。SPSS 所提供的统计功能可以减少诸如此类烦琐的过程。

众数、中位数与算术平均数之间存在一定的关系，这种关系决定于总体分布的状况。当总体分布呈对称的钟型分布时，算术平均数位于分布曲线的对称点上，而该点又是曲线的最高点和中心点，因此，众数、中位数和算术平均数三者相等。当总体分布呈非对称的钟型分布时，由于这三种平均数受极端数值影响程度的不同，因而它们的数值就存在一定的差别，但三者之间仍有一定的关系。当分布右偏时，算术平均数受偏高数值影响较大，其位置必然在众数之右，中位数在众数和算术平均数之间。反之，当次数分布左偏时，算术平均数受偏小数值的影响较大，其位置在众数之左，中位数仍在两者之间。

以上的均值、中位数和众数都是反映数据集中趋势的统计量。

4.1.4　全距（Range）

全距，又称范围，是数据的最大值（Maximum）与最小值（Minimum）之间的绝对差，借以表明总体标志值最大可能的差异范围。全距越长，说明数据越离散；反之，全距越小，说明数据越集中。

用符号表示全距的计算公式为：

$$R = X_{\max} - X_{\min} \tag{4.3}$$

全距的缺点在于其计算方法过于粗略，因为它只考虑总体两端数值的差异，没有考虑中间数值差异的情况，因而它是测定离散程度的一种粗略的方法，不能全面反映总体数据的差异程度。要充分利用每一个数据的信息，就需要利用方差和标准差。

4.1.5　方差（Variance）和标准差（Standard Deviation）

方差是总体所有变量值与其算术平均数偏差平方的平均值，它表示了一组数据分布的离散程度的平均值。标准差是方差的平方根，它表示了一组数据关于平均数的平均离散程度。总体方差：

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{4.4}$$

总体标准差：

$$\sigma = \sqrt{\sigma^2} \tag{4.5}$$

样本方差：

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} \tag{4.6}$$

样本标准差：

$$s=\sqrt{s^2} \tag{4.7}$$

其中，μ为总体平均数，$\bar{x}$为样本平均数，N为总体的个数，n为样本的个数。

虽然标准差有计量单位，而方差无计量单位，但两者的作用相同，故在此仅介绍标准差。标准差用平方的方法消除了正负号，因而它是最常用的、也是最重要的离散趋势统计量。标准差越大，表示变量值之间的差异越大，各数据距离均值越远，则平均数代表性就越低。反之，标准差越小，表示变量值之间的差异越小，各数据距离均值较近，则平均数代表性就越高。

标准差在实际生活中也有广泛的应用。例如，可以用标准差来测定居民收入分配的差异程度，还可以用来反映平均收支、平均结余、平均产量等经济变量的代表性。

全距、方差和标准差都是反映数据离散趋势的统计量。

4.1.6 峰度（Kurtosis）和偏度（Skewness）

峰度是描述总体数据所有取值分布形态陡缓程度的统计量。这个统计量需要与正态分布相比较，峰度为 0 表示该总体数据分布与正态分布的陡缓程度相同；峰度大于 0 表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰；峰度小于 0 表示该总体数据分布与正态分布相比较为平坦，为平顶峰。峰度的绝对值数值越大，表示其分布形态的陡缓程度与正态分布的差异度越大。

峰度的具体计算公式为：

$$Kurtosis=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^4/SD^4-3 \tag{4.8}$$

偏度与峰度类似，它也是描述数据分布形态的统计量，其描述的是某总体取值分布的对称性。这个统计量同样需要与正态分布相比较，偏度为 0 表示其数据分布形态与正态分布的偏斜程度相同；偏度大于 0 表示其数据分布形态与正态分布相比为正偏或右偏，即有一条长尾巴拖在右边，数据右端有较多的极端值；偏度小于 0 表示其数据分布形态与正态分布相比为负偏或左偏，即有一条长尾拖在左边，数据左端有较多的极端值。偏度的绝对值越大，表示其分布形态的偏斜程度越大。

偏度的具体计算公式为：

$$Skewness=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^3/SD^3 \tag{4.9}$$

4.1.7　四分位数（Quartiles）、十分位数（Deciles）和百分位数（Percentiles）

四分位数是将一组数据由小到大（或由大到小）排序后，用 3 个点将全部数据分为四等份，与这 3 个点位置上相对应的数值称为四分位数，分别记为Q_1（第一四分位数）、Q_2（第二四分位数，即中位数）、Q_3（第三四分位数）。其中，Q_3到Q_1之间的距离的一半又称为四分位差，记为Q。四分位差越小，说明中间部分的数据越集中；四分位数越大，则意味着中间部分的数据越分散。

与四分位数类似，十分位数是将一组数据由小到大（或由大到小）排序后，用 9 个点将全部数据分为 10 等份，与这 9 个点位置上相对应的数值称为十分位数，分别记为D_1，D_2，…，D_9，表示 10%的数据落在D_1下，20%的数据落在D_2下，……，90%的数据落在D_9下。

同理，百分位数是将一组数据由小到大（或由大到小）排序后分割为 10 等份，与 99 个分割点位置上相对应的数值称为百分位数，分别记为P_1，P_2，…，P_{99}，表示 1%的数据落在P_1下，2%的数据落在P_2下，……，99%的数据落在P_{99}下。

通过四分位数、十分位数和百分位数，可以大体看出总体数据在哪个区间内更为集中，即它们在一定程度上可以反映出数据的分布情况。

上面的峰度系数、偏度系数和四分位数、十分位数、百分位数，都是反映数据分布状况的统计量。

4.2　数据描述

描述性统计分析是对数据进行基础性的描述。通过得出的数据的平均值（Mean）、和（Sum）、标准差（Std deviation）、最大值（Max）、最小值（Min）、方差（Variance）、全距（range）、平均值标准误（S.E. Mean）、峰度（Kurtosis）、偏度（Skewness）等统计量，来估计原始数据的集中程度、离散状况、分布情况。

操作步骤如下。

01 打开“分析（Analyze）”菜单，选择“描述统计（Descriptive Statistics）”下的“描述（Descriptives）”命令，如图 4.1 所示。

这里，需要提醒的是如果数据文件尚未打开，“分析（Analyze）”菜单下任一功能都不能使用，SPSS 会弹出一个对话框，如图 4.2 所示，提醒用户打开文件。打开文件后，“分析（Analyze）”菜单下的统计功能才能正常使用。

02 选择“描述（Descriptives）”分析后，SPSS 将打开“描述性（Descriptives）”对话框，如图 4.3 所示。



图 4.1 “分析（Analyze）”菜单

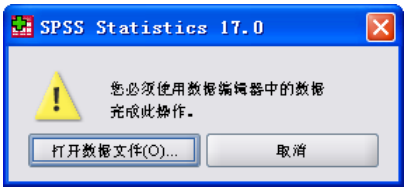


图 4.2 未打开文件提醒对话框



图 4.3 “描述性（Descriptives）”主对话框

在“描述性（Descriptives）”主对话框中，用户可以通过单击从左边源变量中选择一个或者几个变量进入右边的“变量（Variable(s)）”列表框中。

对话框底部有一个“将标准化得分保存为变量（Save standardized values as variables）”复选框，选择该项，将对“变量（Variable(s)）”列表框中被选中变量的数据进行标准化，然后将标准化的结果保存到新变量中。新变量的变量名为原变量的变量名前面添加字母“z”，并被添加在数据编辑窗口中变量的最后一列。

数据标准化的计算公式为：

$$Z = \frac{x - \mu}{\sigma} \tag{4.10}$$

通过标准化，可以将均值为μ、标准差为σ的原变量转化成均值为 0、标准差为 1 的新变量。

“描述性（Descriptives）”主对话框的右端有 5 个按钮，如果还未将左边源变量中的变量添加至“变量（Variable(s)）”列表框中，则“确定（OK）”和“粘贴（Paste）”按钮为灰色显示，不可单击，但“重置（Reset）”、“取消（Cancel）”和“帮助（Help）”按钮可以单击。通过“重置（Reset）”按钮，用户可以将已进入右框的变量全部转移至左框的变量列表中，进行重新选择。

03 单击“选项（Options）”按钮，打开“描述：选项（Descriptives：Options）”对话框，如图 4.4 所示。

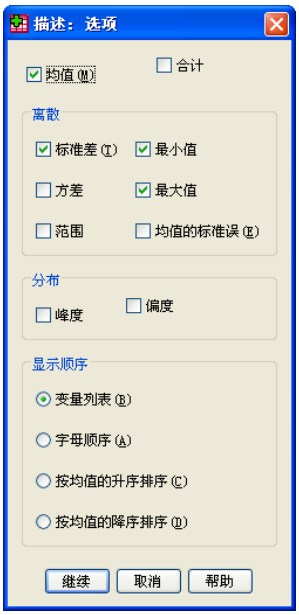


图 4.4　“描述：选项（Descriptives：Options）”对话框

在“描述：选项（Descriptives：Options）”对话框中，用户可以选择所需统计的统计量和图表输出方式。具体对话框中各选项的意义如下。

在对话框中最上面一行是均值（Mean）、合计（Sum）。

离散（Dispersion）栏中的统计量包括：

标准差（Std Deviation）	最小值（Minimum）
方差（Variance）	最大值（Maximum）
范围（Range）	均值的标准误（S.E. mean）

分布（Distribution）栏中的统计量包括：

峰度（Kurtosis）	偏度（Skewness）

显示顺序（Display Order）栏中，用户可以自行选择输出变量的排序方式，包括：

- 变量列表：在结果输出窗口中，用户选择输出的变量将按照变量在数据编辑窗口中原来的排列顺序进行排列。
- 字母顺序（Alphabetic）：在结果输出窗口中，用户选择输出的变量将按照变量名的字母排列顺序进行排列。
- 按均值的升序排序（Ascending Means）：SPSS 将计算每个输出变量的平均值，并按照平均值从小到大对输出变量的顺序进行排列。
- 按均值的降序排序（Descending Means）：SPSS 将计算每个输出变量的平均值，并按照平均值从大到小对输出变量的顺序进行排列。

用户可在“描述：选项（Descriptives：Options）”对话框第一行、离散（Dispersion）栏和分布（Distribution）栏中，选中所需统计的统计量（可多项选择）。SPSS 默认的描述统计量包括均值、标准差、最小值、最大值。在“显示顺序（Display Order）”一栏里，用户只可选择一种变量排序方式，SPSS 的默认选项为“变量列表（Variable List）”。

进行选择后，单击“继续”按钮，即可返回“描述性（Descriptives）”主对话框。

04 单击“确定”按钮，即可在结果输出窗口得到描述性统计分析结果输出表格。

4.3 频数分析

对于一组数据，考察不同的数值出现的频数，或者是数据落入指定区域内的频数，可以了解数据的分布状况。SPSS 17.0 提供了“频数分析”这一功能。通过频数分析，用户在得到描述性统计结果的同时，还能使用户了解变量取值的分布情况，从而使总体数据的分布通过频数分析出的数字得到更为清晰、准确的输出。

频数分析的具体操作步骤如下。

01 打开数据文件，选择“分析（Analyze）”菜单，单击“描述统计（Descriptive Statistics）”下的“频率（Frequencies）”命令。SPSS 将弹出“频率（Frequencies）”主对话框，如图 4.5 所示。

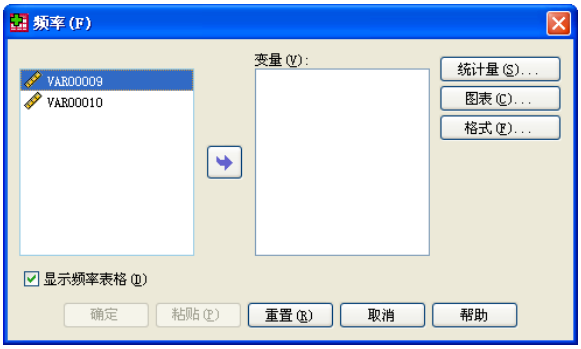


图 4.5 “频率（Frequencies）”主对话框

在“频率（Frequencies）”主对话框中，同样可以通过单击从左边源变量中选择一

个或者几个变量进入右边的“变量（Variable(s)）”列表框中。

对话框底部有一行“显示频率表格（Display frequency tables）”复选框，SPSS 默认选择此项。选择此项后，输出结果将显示频数分布表，否则只显示直方图，不显示频数分布表。

“频率”主对话框的右方有三个按钮，从上到下依次为“统计量（Statistics）”按钮、“图表（Charts）”按钮和“格式（Format）”按钮。单击可进入对应对话框。

02 单击“统计量（Statistics）”按钮，打开“频率：统计量（Frequencies:Statistics）”对话框，如图 4.6 所示。

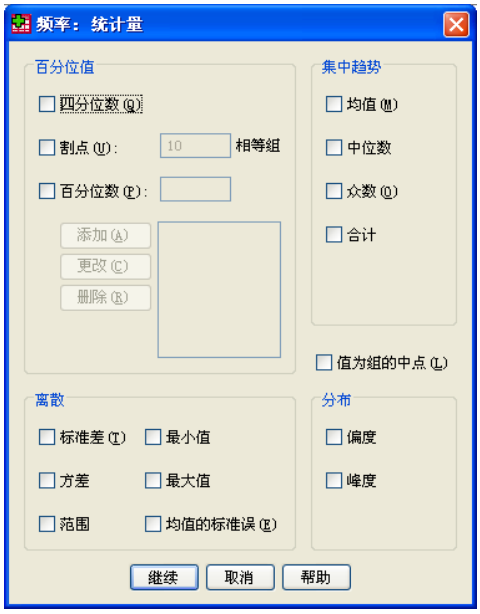


图 4.6　“频率：统计量（Frequencies:Statistics）”对话框

在“频率：统计量（Frequencies:Statistics）”对话框中，用户可以选择所需统计的统计量。具体对话框中各选项的意义如下。

百分位值（Percentile Values）栏：复选项，在此栏中可选择多项。

四分位数（Quartile）。

割点（Cut points）：选择此项，在后面的数值框中输入数值，假设为 N（N 为 2 与 100 之间的整数），则计算并显示 N 分位数。

百分位数（Percentile(s)）：选择此项，在后面的数值框中输入数值，可以有选择地显示百分位数。在数值框中可以输入 0 到 100 之间的整数，输入后，单击“添加（Add）”按钮，将对应的百分位数添加到方框内的列表框中，利用“更改（Change）”按钮和“删除（Remove）”按钮，可以对列表框中的选项进行修改和删除。

离散（Dispersion）栏（复选项）。

标准差（Std Deviation）　　最小值（Minimum）

方差（Variance）　　最大值（Maximum）
范围（Range）　　均值的标准误（S.E. mean）

集中趋势（Central Tendency）栏（复选项）。

均值（Mean）、中位数（Median）、众数（Mode）、合计（Sum）

注意：“集中趋势”栏下方有一个“值为组的中点（Values are group midpoints）”复选框，如果假设数据已经分组，而且数据取值为初始分组的中点，选择此项，将计算百分位数统计和数据的中位数。

分布（Distribution）栏（复选项）。

峰度（Kurtosis）、偏度（Skewness）

用户在“频率：统计量（Frequencies:Statistics）”对话框中单击选中所需统计的统计量后，单击“继续”按钮，即可返回主对话框。

03 单击“图表（Charts）”按钮，打开“频率：图表（Frequencies:Charts）”对话框，如图 4.7 所示。

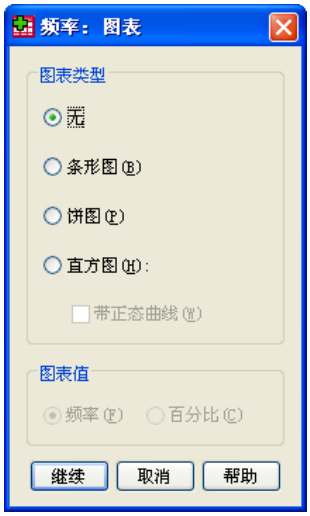


图 4.7 “频率：图表（Frequencies:Charts）”对话框

在“频率：图表（Frequencies:Charts）”对话框中，用户可以选择频数分析的图表。该对话框中各选项的具体意义如下。

图表类型（Chart Type）（单选项）。

无（None）：系统默认选项。

条形图（Bar charts）。

饼图（Pie charts）。

直方图（Histograms）。

注意：如果选择输出“直方图”，可以选择是否在输出的直方图中添加正态分布曲线，如果需要输出正态分布曲线，则可选中“带正态曲线（With normal curve）”复选框。

图表值（Chart Values）（单选项）：可选择图形中分类值的表现形式。

频率（Frequencies）：如果图表类型是直方图，则直方图的纵轴为频数；如果图表类型是饼图，则饼图中每块表示属于该组观测值频数。

百分比（Percentage）：如果图表类型是直方图，则直方图的纵轴为百分比；如果图表类型是饼图，则饼图中每块表示该组的个案数占总数的百分比。

用户在“频率：图表（Frequencies：Charts）”对话框中选择好图表类型和图表分类值后，单击“继续”按钮，即可返回主对话框。

04 单击“格式（Format）”按钮，打开“频率：格式（Frequencies:Format）”对话框，如图 4.8 所示。

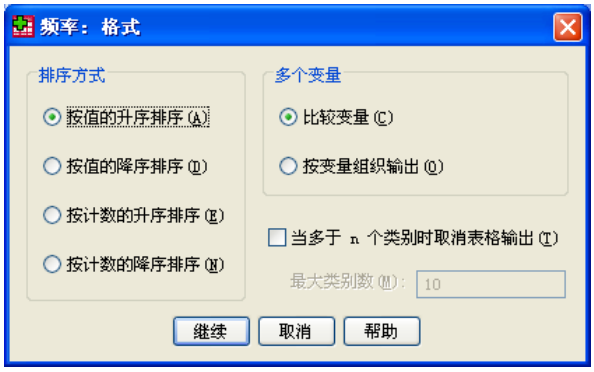


图 4.8　“频率：格式（Frequencies:Format）”对话框

在“频率：格式（Frequencies:Format）”对话框中，用户可以设置频数分布表的输出格式。对话框中各选项的意义如下。

排序方式（Order by）栏（单选项）：，用户可以选择频数分布表中数值及其对应频率的排列顺序。

按值的升序排序（Ascending values）：系统默认选项，频数分布表中将按照数值从小到大排列。

按值的降序排序（Descending values）：频数分布表中将按照数值从大到小排列。

按计数的升序排序（Ascending counts）：频数分布表中将按照频数从小到大排列。

按计数的降序排序（Descending counts）：频数分布表中将按照频数从大到小排列。

注意：如果用户在“频率：图表（Frequencies:Charts）”对话框中选择输出直方图，频数分布表将按照数值顺序排列。

多个变量（Multiple Variables）栏（单选项）：当“频率（Frequencies）”主对话框的“变量（Variable(s)）”列表框中有多个变量时，利用“多个变量（Multiple Variables）”栏可

以设置表格的显示方式。

比较变量（Compare variables）：系统默认选项，SPSS 将所有变量的描述统计的结果显示在同一张表格中，方便用户进行比较分析。

按变量组织输出（Organize output by variable）：SPSS 将对应每个变量，分别输出单独的描述统计表格。

注意：在“频率：格式”对话框的最底端，有一个“当多于*n*个类别时取消表格输出”复选项。通过在“最大类别数数值框中”输入数值，确定频数表输出的方位，即输出数据的组数不得大于窗口中输入的数值。分类数最大参数的默认值是 10。

用户在“频率：格式（Frequencies：Format）”对话框中进行选择后，单击“继续”（Continue）按钮，即可返回“频率（Frequencies）”主对话框。一般情况下，对“频数：格式（Frequencies：Format）”对话框的选项都默认为系统默认值，不做调整。

05 单击“确定”按钮，即可在结果输出窗口中得到频率分布表、描述性统计分析输出表格和用户选择的对应的输出图形。

4.4 探索分析

探索分析是在对数据的基本特征统计量有初步了解的基础上，对数据进行的更为深入、详细的描述性观察分析。它在一般描述性统计指标的基础上，增加有关数据其他特征的文字与图形描述，显得更加细致和全面，有助于用户思考对数据进行进一步分析的方案。主要的分析如下。

观察数据的分布特征：通过绘制箱锁图和茎叶图等图形，直观地反映数据的分布形式和数据的一些规律性，包括考察数据中是否存在异常值等。过大或过小的数据均有可能是奇异值、影响点或错误数据，寻找异常值，并分析原因，然后决定是否从分析中删除这些数据。因为奇异值和影响点往往对分析的影响较大，不能真实地反映数据的总体特征。

正态分布检验：检验数据是否服从正态分布。很多检验能够进行的前提即为总体数据分布服从正态分布。因此，检验数据是否符合正态分布，就决定了它们是否能用只对正态分布数据适用的分析方法。

方差齐性检验：用 Levene 检验比较各组数据的方差是否相等，以判定数据的离散程度是否存在差异。例如在进行独立右边的 T 检验之前，就需要事先确定两组数据的方差是否相同。如果通过分析发现各组数据的方差不同，还需要对数据进行方差分析，那么就需要对数据进行转换使得方差尽可能相同。用 Levene 检验进行方差齐次性检验时，不强求数据必须服从正态分布，它先计算出各个观测值减去组内均值的差，然后再通过这些差值的绝对值进行单因素方差分析。如果得到的显著性水平（Significance）小于 0.05，那么就可以拒绝方差相同的假设。

探索分析的具体操作步骤如下。

01 打开数据文件，选择“分析（Analyze）”菜单，单击“描述统计（Descriptive Statistics）”下的“探索（Explore）”命令，SPSS 将弹出“探索（Explore）”对话框，如图 4.9 所示。

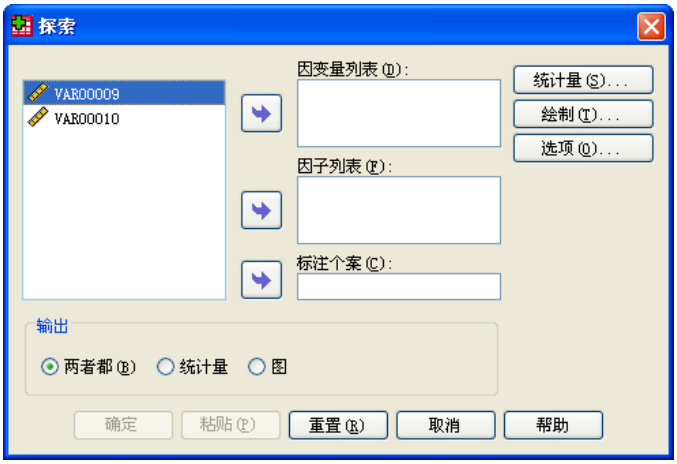


图 4.9 “探索（Explore）”对话框

在“探索（Explore）”对话框中，左边的变量列表为源变量列表，通过单击可选择一个或者几个变量进入右边的“因变量列表（Dependent List）”框、“因子列表（Factor List）”框和“标注个案（Label Cases by）”列表框。因变量是用户所研究的目标变量。因子变量是影响因变量的因素，例如分组变量。标签个案是区分每个个案的变量，例如雇员的 ID 等。例如，研究同一班级男生女生的身高差距时，就可将“身高”变量送入“因变量列表（Dependent List）”框中，将“性别”送入“因子列表（Factor List）”框中，同时将学生的“学号”变量送入“标注个案（Label Cases by）”列表框中。

注意：如果有多个分组变量进入“因子列表（Factor List）”列表框中，那么会以分组变量的各种取值进行组合分组。如两个分组变量各有两种取值，那么输出的结果就会有 4 种组合分组。

在对话框下端的“输出（Display）”框中有三个选项：

- 两者都（Both）：默认选项，表示同时输出描述统计量的统计表格和图形。选择此项将激活右边的“统计量（Statistics）”和“图（Plots）”按钮。
- 统计量（Statistics）：表示只输出统计表格，不输出图表。选择此项将激活右边的“统计量（Statistics）”按钮，“图（Plots）”按钮不被激活。
- 图（Plots）：表示只输出图表，不输出统计表格。选择此项将激活右边的“图（Plots）”按钮，“统计量（Statistics）”按钮不被激活。

02 单击“统计量（Statistics）”按钮，打开“探索：统计量（Explore：Statistics）”对话框，如图 4.10 所示。

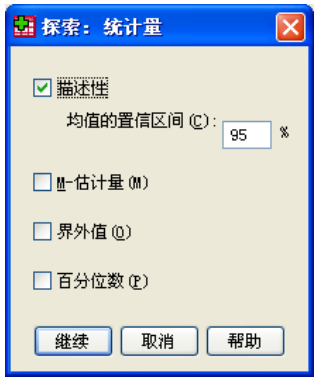


图 4.10 “探索：统计量（Explore：Statistics）”对话框

在“探索：统计量（Explore：Statistics）”对话框中，四个选择项分别如下。

- 描述性（Descriptives）：选择此项，将生成描述性统计表格，表中显示样本数据的描述统计量，包括平均值、中位数、5%调整平均数、标准误差、方差、标准差、最大值、最小值、组距、四分位数、峰度、偏度及峰度和偏度的标准误差。此项为默认选项，在下面的“均值的置信区间（Confidence Interval for Mean）”小窗口中，用户还可输入数值指定均值的置信区间的置信度，系统默认的置信度为 95%。
- M-估计量（M-estimators）：选择此项，将计算并生成稳健估计量。M 估计在计算时对所有个案赋予权重，随个案距分布中心的远近而变化，通过给远离中心值的数据赋予较小的权重来减小异常值的影响。
- 界外值（Outliers）：选择此项，将输出分析数据中的五个最大值和五个最小值作为异常嫌疑值。
- 百分位数（Percentiles）：选择此项，将计算并显示指定的百分位数，包括 5%、10%、25%、50%、75%、90%和 95%等。

“探索：统计（Explore：Statistics）”对话框中的四个选项为复选项，用户可进行多项选择，单击“继续”按钮，即可返回“探索（Explore）”主对话框。

03 单击“绘制（Plots）”按钮，打开“探索：图（Explore：Plots）”对话框，如图 4.11 所示。

“探索：图（Explore：Plots）”对话框中有如下四个选择栏。

（1）箱图（Boxplots）栏（单选项）：箱图，又称箱锁图。如果用户在“探索（Explore）”主对话框的“因变量列表（Dependent List）”框中输入了多个变量名，在此选择栏中进行选择，可确定箱锁图的生成方式。箱锁图中，最底部的水平线段是数据的最小值（异常点除外），顶部的水平线段是数据的最大值（异常点除外），中间矩形箱子的底所在的位置是数据的第一个四分位数（即 25%分位数），箱子顶部所在位置是数据的第三个四分位数据（即 75%分位数）。箱子中间的水平线段刻画的是数据的中位数（即 50%分位数）。

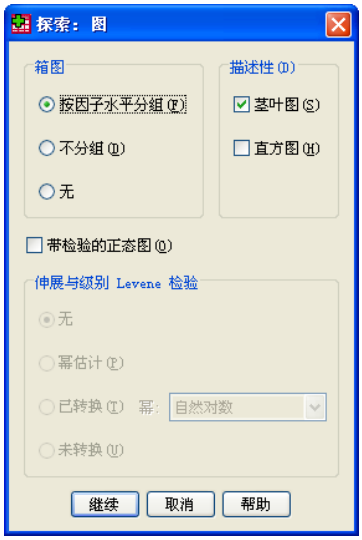


图 4.11　“探索：图（Explore：Plots）”对话框

- 按因子水平分组（Factor levels together，系统默认）：选择此项，将为每个因变量创建一个箱锁图，在每个箱锁图内根据分组变量的不同水平的取值创建箱形单元。
- 不分组（Dependents together）：选择此项，将为每个分组变量的水平创建一个箱锁图，在每个箱锁图内用不同的颜色区分不同因变量所对应的箱形单元，方便用户进行比较。
- 无（None）：选择此项，不创建箱图。

（2）描述性（Descriptive）栏（复选项）：选择该栏内的选项，可以生成茎叶图和（或）直方图。如果箱图（Boxplots）栏内选择的选项不同，则生成的茎叶图和直方图也不相同。选择“按因子水平分组（Factor levels together）”单选钮时，在创建茎叶图和（或）直方图时，首先会根据因变量的不同进行分类，为每一个因变量对应的不同分组变量的不同水平创建一个茎叶图和（或）直方图；选择“不分组（Dependents together）”单选钮时，在创建茎叶图和（或）直方图时，则首先根据不同分组变量水平的不同，为每一个因变量创建一个茎叶图和（或）直方图。

- 茎叶图（Stem-and-leaf，系统默认）：茎叶图主要由三个部分组成，即频率（Frequency）、茎（Stem）和叶（Leaf），在图中按从左到右的顺序依次排列，在图的底端，注明了茎的宽度（Stem Width）和每一叶所代表的个案数（Each Leaf）。茎叶图中，茎表示数据的整数部分，叶表示数据的小数部分（小数位数只有一位，频数的数值有多大，则对应的小数就有多少个），将茎和叶的数值组合起来再乘以茎宽，便是该数据的值。由于茎叶图不仅仅能表示数据的频数分布，还能近似地表示数据的大小，因此它比直方图表达的信息更全面。
- 直方图（Histogram）：直接绘制直方图的步骤详见第 3 章。

（3）带检验的正态图（Normality plots with test，复选项）：选择此项，将进行正态性检验，并生成正态 Q-Q 概率图和无趋势正态 Q-Q 概率图。

（4）伸展与级别 Levene 检验（Spread vs level with Levene Test）栏（单选项）：对所有的展布-水平图进行方差齐次检验和数据转换，同时输出回归直线的斜率以及方差齐性的 Levene's 检验，但如果没有指定分组变量，则此选择项无效。

- 无（None）：不进行 Levene 检验，系统默认。选择此项，SPSS 将不产生回归直线的斜率和方差齐性检验。
- 幂估计（Power Estimation）：对每组数据产生一个中位数的自然对数及四个分位数的自然对数的散点图。
- 已转换（Transformed）：变换原始数据，用户可在后面的参数框中选择数据变换类型。
- 未转换（Untransformed）：不变换原始数据时选择此项。

用户在“探索：绘图（Explore：Plots）”对话框进行选择后，单击“继续（Continue）”按钮，即可返回“探索（Explore）”主对话框。

04 单击“选项（Options）”按钮，打开“探索：选项（Explore：Options）”对话框，如图 4.12 所示。

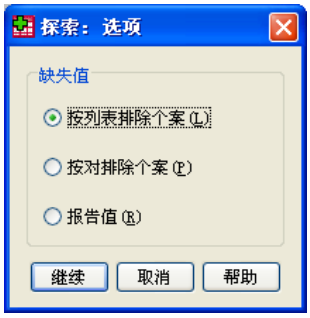


图 4.12 “探索：选项（Explore：Options）”对话框

在“探索：选项（Explore：Options）”对话框中，可选择缺失值的处理方式，SPSS 17.0 提供了三种处理方式：

- 按列表排除个案（Exclude cases listwise，系统默认）：选择此项，对所有的分析过程剔除分组变量和因变量中所有带有缺失值的个案数据；
- 按对排除个案（Exclude cases pairwise）：同时剔除带缺失值的个案及与缺失值有成对关系的个案。在当前分析过程用到的变量数据中剔除带有缺失值的个案数据，在其他分析过程中可能包含缺失值；
- 报告值（Report values）：选择此项，将分组变量的缺失值单独分为一组，在输出频数表的同时输出缺失值。

用户在“探索：选项（Explore：Options）”对话框进行选择后，单击“继续（Continue）”按钮，即可返回“探索（Explore）”主对话框。

05 单击“确定（OK）”按钮，即可在结果输出窗口得到探索分析过程的数据概述、基本统计描述表、极端值列表、正态分布检验、方差齐次检验、茎叶图、直方图、箱锁图、正态分布 Q-Q 图、离散正态分布 Q-Q 图等图表。

4.5　交叉列联表分析

在实际分析中，除了需要对单个变量的数据分布情况进行分析外，还需要掌握多个变量在不同取值情况下的数据分布情况，从而进一步深入分析变量之间的相互影响和关系，这种分析就称为交叉列联表分析。

当所观察的现象同时与两个因素有关时，如某种服装的销量受价格和居民收入的影响，某种产品的生产成本受原材料价格和产量的影响等，通过交叉列联表分析，可以较好地反映出这两个因素之间有无关联性以及两个因素与所观察现象之间的相关关系。

因此，数据交叉列联表分析主要包括两个基本任务：一是根据收集的样本数据，产生二维或多维交叉列联表；二是在交叉列联表的基础上，对两两变量间是否存在相关性进行检验。要获得变量之间的相关性，仅仅靠描述性统计的数据是不够的，还需要借助一些表示变量间相关程度的统计量和一些非参数检验的方法。

常用的衡量变量间相关程度的统计量是简单相关系数，但在交叉列联表分析中，由于行列变量往往不是连续变量，不符合计算简单相关系数的前提条件。因此，需要根据变量的性质，选择其他的相关系数，如 Kendall 等级相关系数、Eta 值等。

SPSS 17.0 提供了多种适用于不同类型数据的相关系数表达，这些相关性检验的零假设都是：行和列变量之间相互独立，不存在显著的相关关系。根据 SPSS 检验后得出的相伴概率（Significance）判断是否存在相关关系。如果相伴概率小于显著性水平 0.05，那么拒绝零假设，行列变量之间彼此相关；如果相伴概率大于显著性水平 0.05，那么接受原假设，行列变量之间彼此独立。

在交叉列联表分析中，SPSS 所提供的相关关系的检验方法主要有以下三种：

- **卡方统计检验：**常用于检验行列变量之间是否相关。计算公式为：

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{4.11}$$

其中，f_0表示实际观察频数，f_e表示期望频数。

卡方统计量服从（行数–1）×（列数–1）个自由度的卡方统计。SPSS 在计算卡方统计量时，同时给出相应的相伴概率，由此判断行列变量之间是否相关。

- **列联系数**（Contingency coefficient）：常用于名义变量之间的相关系数计算。计算公式由卡方统计量修改而得，公式如下。

$$C = \sqrt{\frac{x^2}{x^2 + N}} \tag{4.12}$$

- **ψ系数（Phi and Cramer's V）**：常用于名义变量之间的相关系数计算。计算公式由卡方统计量修改而得，公式如下。

$$V=\sqrt{\frac{x^2}{N(K-1)}} \tag{4.13}$$

ψ系数介于 0 和 1 之间，其中，K 为行数和列数较小的实际数。

交叉列联表分析的具体操作步骤如下。

01 打开数据文件，选择“分析（Analyze）”菜单，单击“描述统计（Descriptive Statistics）”下的“交叉表（Crosstabs）”命令。“交叉表（Crosstabs）”主对话框如图 4.13 所示。



图 4.13 “交叉表（Crosstabs）”对话框

在“交叉表（Crosstabs）”对话框中，左边的变量列表为源变量列表，通过单击可选择一个或者几个变量进入右边的“行（Row(s)）”变量列表框、“列（Column(s)）”变量列表框和“层（Layer）”变量列表框中。

如果是二维列联表分析，只需选择行列变量即可。但如进行三维以上的列联表，可以将其他变量作为控制变量选到“层（Layer）”变量列表框中。有多个层控制变量时，可以根据实际的分析要求确定它们的层次，既可以是同层次的也可以是逐层叠加的。

在“交叉列联表分析”对话框底端，有两个可选择项：

- 显示复式条形图（Display clustered bar chart）：指定绘制各个变量不同交叉取值下关于频数分布的柱形图；
- 取消表格（Suppress table）：不输出列联表具体表格，而直接显示交叉列联表分析过程中的统计量，如果没有选中统计量，则不产生任何结果。所以，一般情况下，只有在分析行列变量间关系时选择此项。

“交叉表（Crosstabs）”对话框的右端有四个按钮，从上到下依次为“精确（Exact）”

按钮、“统计量（Statistics）”按钮、“单元格（Cells）”按钮、“格式（Format）”按钮。单击可进入相应对话框。

02 单击“精确（Exact）”按钮，打开“精确检验（Exact Tests）”对话框，如图 4.14 所示。

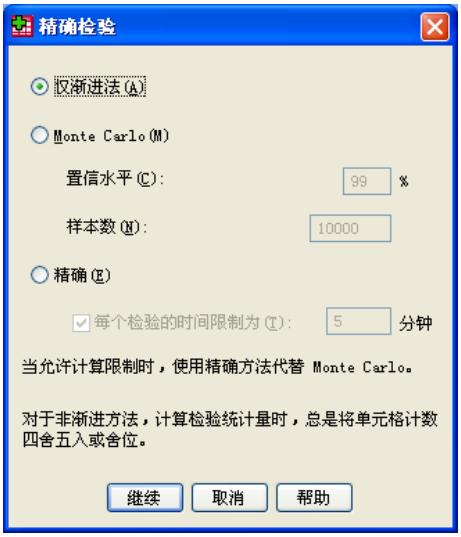


图 4.14 “精确检验（Exact Tests）”对话框

“精确检验（Exact Tests）”对话框提供了三种用于不同条件的检验方式来检验行列变量的相关性。用户可选择以下三种检验方式中的一种；

- 仅渐进法（Asymptotic only）：适用于具有渐近分布的大样本数据，SPSS 默认选择该项；
- Monte Carlo（蒙特卡罗法）：此项为精确显著性水平值的无偏估计，无需数据具有渐近分布的假设，是一种非常有效的计算确切显著性水平的方法。在“置信水平（Confidence Level）”参数框内输入数据，可以确定置信区间的大小，一般为 90，95，99。在“样本数（Number of samples）”参数框中可以输入数据的样本容量；
- 精确（Exact）：观察结果概率，同时在下面的“每个检验的时间限制为（Time limit per test）”的参数框内，选择进行精确检验的最大时间限度。

用户在“精确检验（Exact Tests）”对话框内进行选择后，单击“继续（Continue）”按钮，即可返回“交叉表（Crosstabs）”主对话框。一般情况下，对“精确检验（Exact Tests）”对话框的选项都默认为系统默认值，不做调整。

03 单击“统计量（Statistics）”按钮，打开“交叉表：统计量（Crosstabs：Statistics）”对话框，如图 4.15 所示。

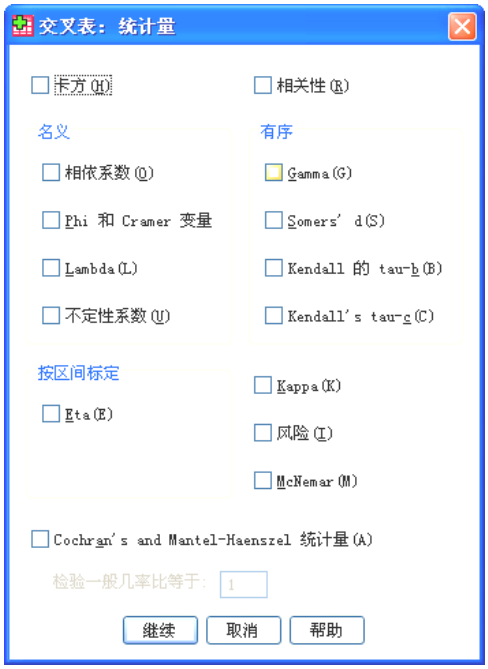


图 4.15 “交叉表：统计（Crosstabs：Statistics）”对话框

在“交叉表：统计量（Crosstabs：Statistics）”对话框中，用户可以选择输出合适的统计检验统计量。对话框中各选项的意义如下。

- 卡方(Chi-square)检验复选项：检验列联表行列变量的独立性检验，也被称为Pearson chi-square检验、x^2检验。
- 相关性（Correlations）检验复选项：输出列联表行列变量的 Pearson 相关系数或 Spearman 相关系数。
- 名义（Nominal）栏：适用于名称变量统计量。

 相依系数（Contingency coefficient）：表征变量之间的相关性弱，该系数越大，变量之间的相关性越强，该系数越小，变量之间的相关性越小。

 Phi 和 Cramer 变量（ψ系数）：反映变量之间的 Phi 相关性。

 Lambda（λ系数）：在自变量预测中用于反映比例缩减误差，其值为 1 时表明自变量预测应变量好，为 0 时表明自变量预测应变量差。

不定性系数（Uncertainty coefficient）：以熵为标准的比例缩减误差，其值接近 1 时表明后一变量的信息很大程度来自前一变量，其值接近 0 时表明后一变量的信息与前一变量无关。

- 有序（Ordinal）栏：适用于有序变量的统计量。

 Gamma 系数（γ系数）：两个有序变量之间的关联性的对称检验。其数值界于 0~1 之间，当所有观察实际数集中于左上角和右下角时，取值为 1，表示两个变量之

间有很强的相关，取值为 0 时，表示两个变量之间相互独立。

Somers'd 值：两个有序变量之间的关联性的检验，取值范围为[-1，1]。

Kendall 的 tau-b 值：考虑有结的秩或等级变量关联性的非参数检验，相同的观察值选入计算过程中，取值范围为[-1，1]。

Kendall′s tau-c 值：忽略有结的秩或等级变量关联性的非参数检验，相同的观察值不选入计算过程，取值范围为[-1，1]。

- 按区间标定（Nominal by interval）栏：适用于一个名义变量和一个等距变量的相关性检验。
- Kappa 系数：检验数据内部的一致性，仅适用于具有相同分类值和相同分类数量的变量交叉表。
- Eta 值：其平方值可认为是应变量受不同因素影响所致方差的比例。
- 风险（相对危险度）：检验事件发生和某因素之间的关联性。
- McNemar 检验：主要用于检验配对的资料率（相当于配对卡方检验）。
- Cochran's and Mantel-Haenszel 统计量（Cochran's and Mantel-Haenszel）复选框：适用于在一个二值因素变量和一个二值响应变量之间的独立性检验。

用户在“交叉表：统计（Crosstabs：Statistics）”对话框内进行选择后，单击“继续（Continue）”按钮，即可返回“交叉表（Crosstabs）”主对话框。一般情况下，不对“交叉表：统计（Crosstabs：Statistics）”对话框内的选项做选择或选择较为常用的卡方检验。

04 单击“单元格（Cells）”按钮，打开“交叉表：单元显示（Crosstabs：Cell Display）”对话框，如图 4.16 所示。

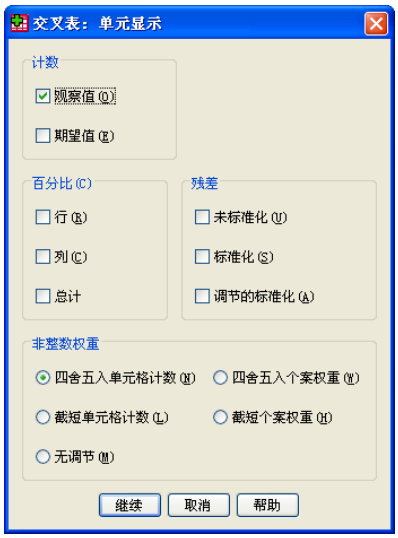


图 4.16　“交叉表：单元显示（Crosstabs：Cell Display）”对话框

在“交叉表：单元显示（Crosstabs：Cell Display）”对话框中，用户可以指定列联表单元

格中的输出内容。SPSS 17.0 默认在交叉列联表中输出实际的观察值，但观察数有时不能确切地反映事物的实质，因此还需要输出其他的数据项。对话框中各选项的具体含义如下。

- 计数（Counts）栏

 观察值（Observed）：系统默认选项，表示输出为实际观察值。

 期望值（Expected）：表示输出为理论值。

- 百分比（Percentages）栏

 行（Row）百分比：以行为单元，统计行变量的百分比。

 列（Column）百分比：以列为单元，统计列变量的百分比。

 总计（Total）百分比：行列变量的百分比都进行输出。

- 残差（Residuals）栏

 未标准化（Unstandardized）：输出非标准化残差，为实际数与理论数的差值。

 标准化（Standardized）：输出标准化残差，为实际数与理论数的差值除理论数。

 调节的标准化（Adjusted standardized）：输出修正标准化残差，为标准误确定的单元格残差。

- 非整数权重（Noninteger Weights）栏

 四舍五入单元格计数（Round cell counts，系统默认）：将单元格计数的非整数部分的尾数四舍五入为整数。

 四舍五入个案权重（Truncate cell counts）：将单元格计数的非整数部分的尾数舍去，直接化为整数。

 截短单元格计数（Round case Weights）：将个案权数的非整数部分的尾数四舍五入为整数。

 截短个案权重（Truncate case Weights）：将个案权数的非整数部分的尾数四舍五入为整数。

 无调节（No adjustments）：不对计数数据进行调整。

用户在“交叉表：单元显示（Crosstabs：Cell Display）”对话框内进行选择后，单击“继续（Continue）”，即可返回“交叉表（Crosstabs）”主对话框。一般情况下，对“交叉表：单元显示（Crosstabs：Cell Display）”对话框的选项都默认为系统默认值，不做调整。

05 单击“格式（Format）”按钮，打开“交叉表：表格格式（Crosstabs：Table Format）”对话框，如图 4.17 所示。

在“交叉表：表格格式（Crosstabs：Table Format）”对话框中，用户可以指定列联表的输出排列顺序。对话框中各选项的具体含义如下。

在行序（Row Order）栏有如下两个选项：

- 升序（Ascending）：系统默认，以升序显示各变量值；
- 降序（Descending）：以降序显示各变量值。

用户在“交叉表：表格格式（Crosstabs：Table Format）”对话框内进行选择后，单击“继续（Continue）”按钮，即可返回“交叉表（Crosstabs）”主对话框。

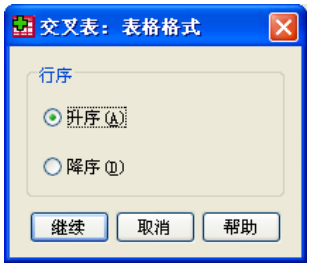


图 4.17 “交叉表：表格格式（Crosstabs：Table Format）”对话框

06 在“交叉表（Crosstabs）”对话框中单击“确定（OK）”按钮，可在输出窗口得到数据概述、交叉列联表（Crosstabulation）、卡方检验表、交叉分组下频数分布柱形图、相对危险性估计等图表。

4.6　比率分析

比率分析主要用于对两变量间变量值的比率变化进行描述分析，适用于定距型变量（scale）。比如，可以对受处分的学生和没受处分的学生的上网时间与用计算机的时间进行比率分析。

比率分析生成比率变量，并对该比率变量计算基本描述性统计量（如均值、中位数、标准差、全距等），进而刻画出比率变量的集中趋势和离散程度。除此之外，SPSS 17.0 还提供了其他相对比描述指标，大致也属于集中趋势描述指标和离散程度描述指标的范畴。

01 打开数据文件，选择“分析（Analyze）”菜单，单击“描述统计（Descriptive Statistics）”下的“比率（Ratio）”按钮。打开“比值统计量（Ratio Statistics）”主对话框，如图 4.18 所示。



图 4.18 “比值统计量（Ratio Statistics）”主对话框

在“比值统计量（Ratio Statistics）”主对话框中，左边的变量列表为源变量列表，通过单击可选择一个或者几个变量进入右边的“分子（Numerator）”变量列表框、“分母（Denominator）”变量列表框和“组变量（Group Variable）”列表框中。做不同组间的比值比较时，需要将变量分组，将变量选择到“组变量（Group Variable）”列表框中后，可以选择将变量按照分组变量进行升序或者降序排列。

“比值统计量”对话框的左下角有如下两个选项：

- 显示结果（Display results，系统默认）；
- 将结果保存到外部文件（Save results to external file）：如果选择该项，“文件（File）”按钮即取消灰色显示，单击“文件（File）”按钮可将输出结果保存至用户指定位置。

“比值统计量（Ratio Statistics）”对话框的右下角是“统计量（Statistics）”按钮，单击此按钮将打开“比率统计量：统计（Ratio Statistics：Statistics）”对话框。

02 单击“统计量（Statistics）”按钮，打开“比率统计量：统计量（Ratio Statistics：Statistics）”对话框，如图 4.19 所示。

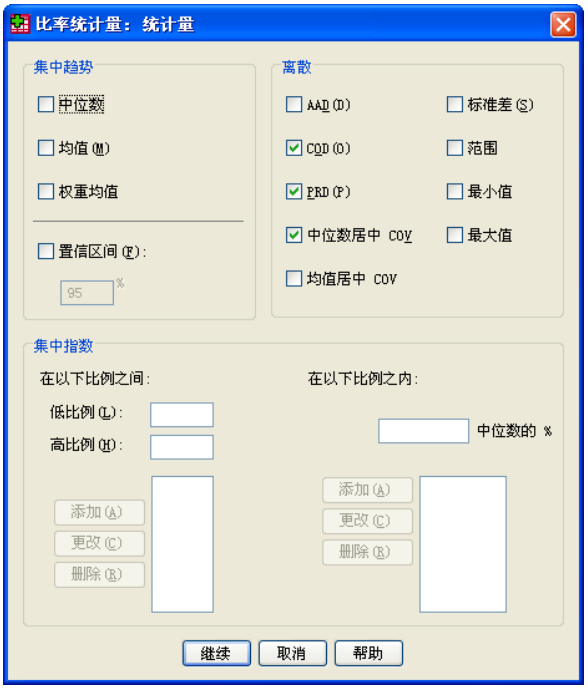


图 4.19 “比率统计量：统计（Ratio Statistics：Statistics）”对话框

在“比率统计量：统计量（Ratio Statistics：Statistics）”对话框中，用户可以选择对比率变量计算基本描述性统计量（如均值、中位数、标准差、全距等），并描述比率变量的集中趋势和离散程度。对话框中各选项的具体含义如下：

- 集中趋势（Central Tendency）栏（复选项）

中位数（Median）

均值（Mean）

权重均值（Weighted Mean）：指两变量均值的比。

置信区间（Confidence intervals）：默认置信区间的置信度为 95%。

- 离散（Dispersion）栏（复选项）

AAD（Average Absolute Deviation，平均绝对离差）：其计算公式为：

$$AAD = \frac{\sum | R_i - M |}{N} \tag{4.14}$$

其中，Ri 是比率数，M 是比率变量的中位数，N 为样本数。

COD（Coefficient of Dispersion，离散系数）：其计算公式为：

$$COD = \frac{\dfrac{\sum | R_i - \bar{R} |}{N}}{M} \tag{4.15}$$

PRD（Price-related Differential，相关价格微分）：是比率均值与加权比率均值的比。

均值居中的 COV（Mean Centered COV）：通常意义下的变异系数，即标准差除以均值。

中位数据中的 COV（Median Centered COV）：计算公式为：

$$COV = \frac{\dfrac{\sum (R_i - M)^2}{N}}{M} \tag{4.16}$$

标准差（Standard deviation）

范围（Range）

最小值（Minimum）

最大值（Maximum）

- 集中指数（Concentration Index）栏

在以下比例之间（Ratios Between）：通过在低比例（Low Proportion）和高比例（High Proportion）后面的窗口中输入数值，可以控制比率的范围。输入后，单击“增加（Add）”按钮，将对应的比率范围添加到方框内的列表框中，利用“更改（Change）”按钮和“删除（Remove）”按钮，可以对列表框中的选项进行修改和删除。

- 在以下比例之内（Ratios Within）：通过在中位数的%（% of median）前面的窗口中输入数值，可以控制比率相对于中位数的范围。输入后，单击“增加（Add）”按钮，将对应的比率限制条件添加到方框内的列表框中，利用“更改（Change）”按钮和“删除（Remove）”按钮，可以对列表框中的选项进行修改和删除。

用户在“比率统计量：统计（Ratio Statistics：Statistics）”对话框内进行选择后，单击“继续（Continue）”按钮，即可返回“比率统计量（Ratio Statistics）”主对话框。

03 在“比率统计量（Ratio Statistics）”主对话框中单击“完成”按钮，可在输出窗口得到分组描述结果、比率分析结果等图表。

4.7 P-P 图、Q-Q 图

P-P 图是根据变量的累积比例与指定分布的累积比例之间的关系所绘制的图形。通过 P-P 图可以检验数据是否符合指定的分布。当数据符合指定分布时，P-P 图中各点近似成一条直线。如果 P-P 图中各点不成直线，但有一定规律，可以对变量数据进行转换，使转换后的数据更接近指定分布。

Q-Q 图同样可以用于检验数据的分布，所不同的是，Q-Q 图是用变量数据分布的分位数与所指定分布的分位数之间的关系曲线来进行检验的。

由于 P-P 图和 Q-Q 图的用途完全相同，只是检验方法存在差异，SPSS 17.0 中用于作出 P-P 图的对话框和用于作出 Q-Q 图的对话框完全一致，下面将对两者统一加以说明。

具体操作步骤如下。

打开数据文件，选择“分析（Analyze）”菜单，单击“描述性统计（Descriptive Statistics）”下的“P-P 图（P-P Plots）”或“Q-Q 图（Q-Q Plots）”命令。“P-P 图（P-P Plots）”和“Q-Q 图（Q-Q Plots）”对话框分别如图 4.20 和图 4.21 所示。

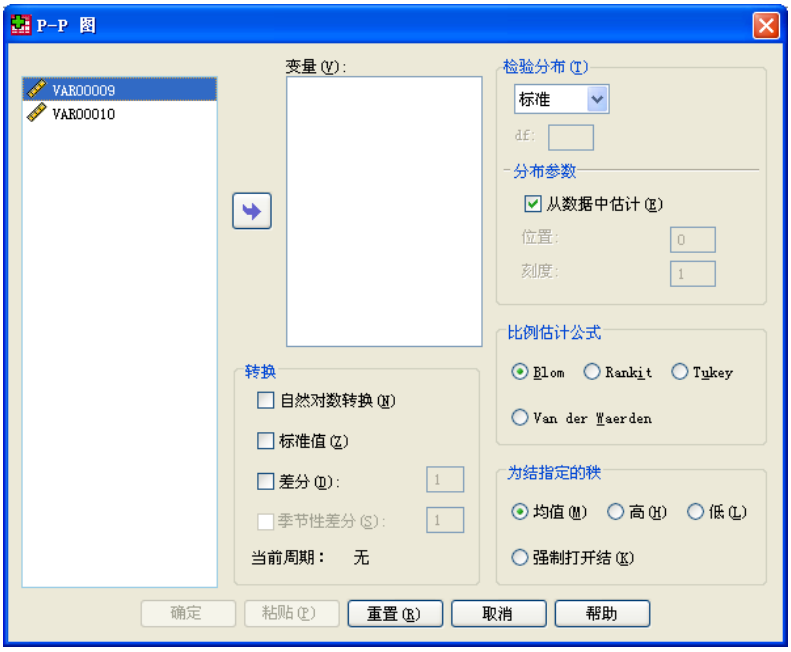


图 4.20 “P-P 图（P-P Plots）”对话框

在“P-P 图（P-P Plots）”或“Q-Q 图（Q-Q Plots）”对话框中，最左边的变量列表为源变量列表，通过单击 可选择一个或者几个变量进入位于对话框中间的“变量（Variables）”列表框中。根据这些变量数据可创建 P-P 图或 Q-Q 图，并进行分布检验。

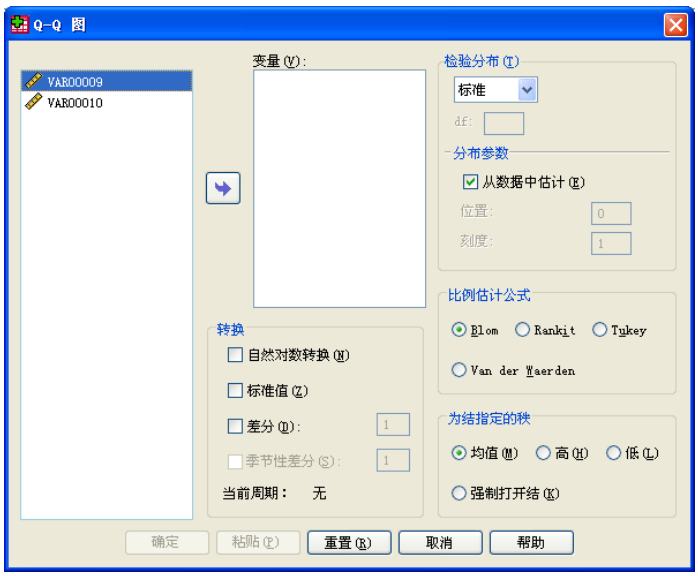


图 4.21　“Q-Q 图（Q-Q Plots）”对话框

“P-P 图（P-P Plots）”或“Q-Q 图（Q-Q Plots）”对话框的中下方和右方有五个选项栏，选项栏中各选项的意义如下。

- 转换（Transform）栏（复选项）

自然对数转换（Natural log transform）：选择此项，对当前变量的数据取自然对数，即将原有变量转换成以自然数 e 为底的对数变量。

标准值（Standardize values）：选择此项，将当前变量的数据转换为标准值，即转换后变量数据的均值为 0，方差为 1。

差分（Difference）：选择此项，对当前变量的数据进行差分转换，即利用变量中连续数据之间的差值来转换数据。选择此项后，后面的窗口将变为可用，在其中输入一个正整数，以确定转换的差分度，默认值为 1。

季节性差分（Seasonally difference）：用于确定指明计算时间序列的季节差分。只有在对当前变量的数据序列定义了周期（通过主菜单中的“数据（Data）”菜单中的“定义日期（Define Dates）”选项定义）以后才可用，如果当前周期为 0，将不能计算季节差分。选择此项，在后面的窗口中输入正整数，然后根据该正整数所确定的范围来计算该范围内数据的差值，并用该差值来转换原数据。

需要注意的是，这些数据转换并不改变变量中的变量值，只影响正态概率图。

- 检验分布（Test Distribution）栏：可选择不同的分布类型，检验变量数据是否符合所选分布。单击选项分布栏下的箭头可选择不同的分布类型。SPSS 默认的检验分布是正态分布（Normal）。

P-P 图（或 Q-Q 图）可检验的分布包括：

贝塔分布（Beta）　　T 分布（Student t）

卡方分布（Chi-square）　　伽玛分布（Gamma）
指数分布（Exponential）　　半正态分布（Half-normal）
逻辑斯诺分布（Logistic）　　拉普拉斯分布（Laplace）
对数正态分布（Lognormal）　　威布尔分布（Weibull）
正态分布（Normal）　　均匀分布（Uniform）
帕累托分布（Pareto）　　dt：在该窗口中输入正整数，表示所选分布的自由度

- 分布参数（Distribution Parameters）栏：在该栏内可输入所选分布类型的参数，选择的分布不同，参数输入窗口也不同。

从数据估计（Estimate from data）：为系统默认选项，选择此项，系统将自动从数据中推测数据分布的参数，否则就要在该选项下方的参数框中根据需要自行指定。

Location（位置）参数窗口：选择正态分布时，用户自行输入位置参数。

Scale（比例）参数窗口：选择正态分布时，用户自行输入比例参数。

- 比例估计公式（Proportion Estimation Formula）栏（单选栏）

Blom 方法：计算公式为 $\frac{r-3/8}{n+1/4}$　　（4.17）

Rankit 方法：计算公式为 $\frac{r-1/2}{n}$　　（4.18）

Turkey 方法：计算公式为 $\frac{r-1/3}{n+1/3}$　　（419）

Van der Waerden 方法：计算公式为 $\frac{r}{n+1}$　　（4.20）

以上公式中，n 表示个案的数目，r 是从 1 到 n 的秩次。

- 为结指定的秩（Rank Assigned to Ties）栏（单选栏）

均值（Mean）：用连接值的平均秩序指定顺序。

高（High）：用连接值的最大秩指定顺序。

低（High）：用连接值的最小秩指定顺序。

强制打开结（Break ties arbitrarily）：忽略个案权重的影响。

4.8　本章小结

本章介绍了“分析菜单”下，“描述性统计”命令的各种功能的用途和用法。通过本章对描述性统计基础方法的学习，读者可以在数据的基础上，运用描述原始数据集中程度、离散状况、分布情况的统计量，对数据的总体特征进行归纳。

第二篇

SPSS 统计分析常用模块篇

- 第 5 章　均值分析与方差分析
- 第 6 章　参数检验
- 第 7 章　非参数检验
- 第 8 章　相关分析
- 第 9 章　回归分析
- 第 10 章　聚类分析
- 第 11 章　判别分析
- 第 12 章　因子分析与主成分分析
- 第 13 章　可靠性分析
- 第 14 章　时间序列分析
- 第 15 章　生存分析

第5章 均值分析与方差分析

本章主要介绍均值分析和方差分析。均值分析可以按某数值或定位变量分组，求出各组的统计量。在统计分析中采用抽样方法时，会使样本统计数与总体参数之间存在差异，均值比较分析可以推断样本均值之间，或样本均值与总体均值之间的差异是否具有显著性意义。方差表示统计数据的离散程度，包括单因素方差分析、多因素方差分析、重复度量和协方差分析。均值分析和方差分析是专业统计分析的重要基础，在经济、医学等领域有广泛的应用。

在 SPSS 中进行方差分析时，可以利用“分析→均值比较”命令实现均值分析和单因素 ANOVA，菜单内容如图 5.1 所示。利用“分析→一般线性模型”命令实现单变量、多因变量方差分析、重复度量和方差成分分析，菜单内容如图 5.2 所示。

本章学习要点：

- 均值分析；
- 方差分析概述；
- 单因素方差分析；
- 多因素方差分析；
- 重复度量；
- 协方差分析。

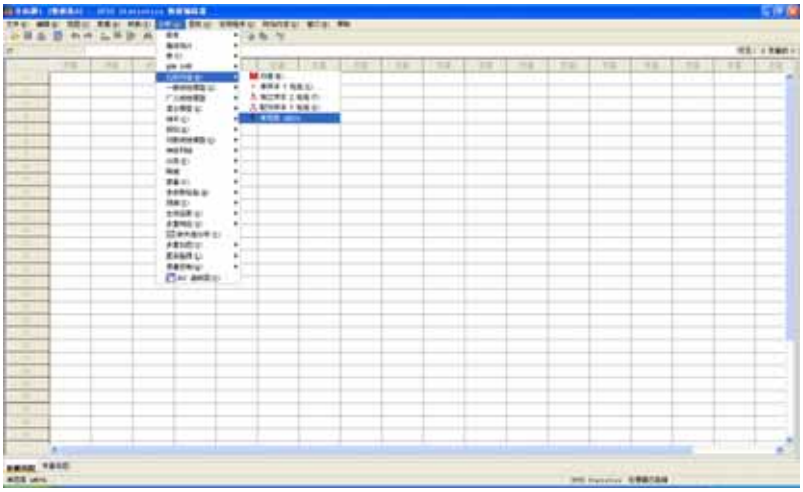


图 5.1　SPSS 单因素 ANOVA 模块

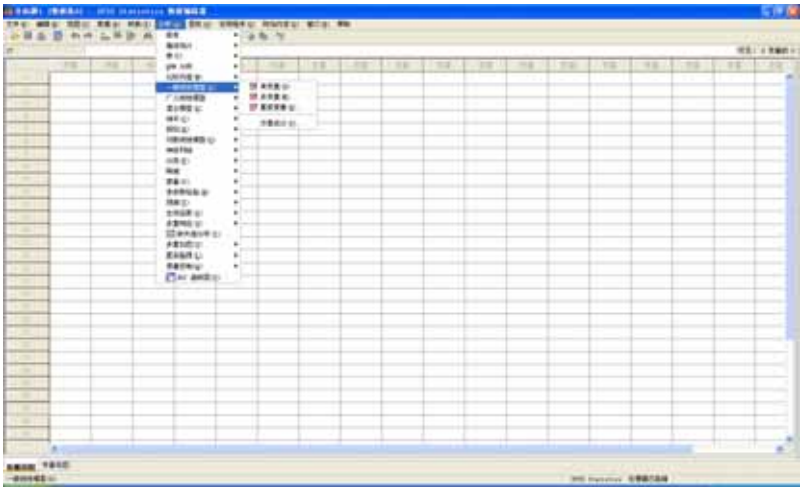


图 5.2　SPSS 其他方差分析

5.1　均值分析

均值分析过程的优势在于所有的描述性统计量均按因变量的取值分组计算，无需像其他程序那样必须先运行文件拆分过程。在输出结果中各组的描述指针被放在一起，也便于相互比较。如果需要，均值分析过程还可以直接输出方差分析结果，并计算相应的相关性指针而无需再次运行其他过程，显然要比先描述再比较要方便得多。

SPSS 的“均值比较”模块提供了均值分析的过程，下面介绍基本操作和各个对话框的功能。

SPSS 均值分析具体步骤如下。

- 打开“ 均值分析 ”对话框

依次单击菜单“ 分析→均值比较→均值 ”，进入“ 均值 ”对话框，如图 5.3 所示。



图 5.3 “ 均值 ”对话框

- 选入变量

在变量列表框中选择因变量，单击 按钮，将其选入到右侧的“ 因变量列表 ”列表框，可以选择一个以上的变量；在变量列表框中选择因子变量，单击 按钮，将其移动到“ 自变量列表 ”列表框。

- “ 选项 ”设置

单击如图 5.3 所示对话框中的“ 选项 ”按钮，打开“ 均值：选项 ”对话框，如图 5.4 所示。



图 5.4 “ 均值：选项 ”对话框

左侧列表框中为可选的描述性统计量，共包含 18 项，分别为中位数、组内中位数（Grouped Median）、均值的标准误（Std.Error of Mean）、合计、最小值、最大值、范围、第一个（First）、最后一个（Last）、方差、峰度（Kurtosis）、峰度的标准误（Std.Error of Kurtosis）、偏度（Skewness）、偏度的标准误（Std.Error of Skewness）、调和均值（Harmonic Mean）、

几何平均数（Geometric Mean）、总和的百分比（Percent of Total Sum）和占总样本计数的百分比（Percent of Total N）。

右侧列表框表示选入的描述性统计量，默认情况下已经选入了均值、个案数（Number of Cases）和标准差（Standard Deviation）三种统计量。

“ 第一层的统计量 ” 选项组用于选择检验第一层（注意：是第一层而不是第一个）的分组变量对结果变量的影响是否有统计学意义。包含两个选项，其意义如表 5.1 所示。

表 5.1　第一层的统计量选项组选项及其意义

选　项	意　义
方差分析表和 eta 值	对分组变量进行单因素方差分析，并计算 eta 值。它用于度量分组变量和结果变量间的方差分析，eta 的平方表示由组间差异所解释的结果变量的方差的比例，即 SS 组间/SS 汇总
线性检验	即不同组的平均数间是否存在线性趋势。实际上也就是进行单因素方差分析的同时，计算出 *R* 和 *R2*，前者是用户熟悉的相关系数，后者即决定系数。注意，如果第一层中有多个变量，则 SPSS 只对最后一个因变量计算 *R* 和 *R2*

● 执行操作

在 “ 均值 ” 主对话框中，单击 “ 确定 ” 按钮，执行操作。

【示例 5.1】均值分析示例。

在治疗肝癌的药物研究中，为了提高治疗药物在靶器官——肝脏的浓度，降低在非靶器官如心脏的浓度，进行 2×3×2 析因设计，即设置 3 个因素，第一个因素是药物（drug），有两个水平，分别是 “ 丝裂霉素+高分子物质+磁性物质 ”（实验组=1）和 “ 丝裂霉素 ”（实验组=2）；第二个因素是时间（time），有 3 个水平，分别为给药后 15min（=1）、30min（=2）和 60min（=3）；第三个因素是器官（organ），有两个水平，分别为心脏（=1）和肝脏（=2）。将 60 只小白鼠随机分为 12 组（即 2×3×2 种组合），每组 5 只，即重复例数为 5。观察指标（反应变量）为组织中丝裂霉素的浓度（$\mu g/g$），结果如表 5.2 所示。

表 5.2　小白鼠不同器官组织的丝裂霉素浓度（$\mu g/g$）检测结果

	药物实验组 1（drug1）			药物实验组 2（drug2）		
	15 分	30 分	60 分	15 分	30 分	60 分
心脏（organ1）	0.1189	0.3498	0.2404	0.3482	0.6204	0.3968
	0.1236	0.3227	0.2676	0.3646	0.6544	0.3935
	0.1333	0.3488	0.2505	0.3780	0.6779	0.3942
	0.1031	0.3119	0.2642	0.3562	0.6312	0.3770
	0.0920	0.3270	0.2434	0.3596	0.6221	0.3918
肝脏（organ2）	0.7787	3.6153	0.5643	0.1613	0.3774	0.1194
	0.7798	3.4654	0.5691	0.1663	0.3566	0.0929
	0.7560	3.4980	0.5799	0.1502	0.3748	0.1050
	0.7745	3.3174	0.5859	0.1124	0.3829	0.0985
	0.7999	3.3617	0.5628	0.1637	0.3942	0.1196

本例具体的操作步骤如下。

01 打开数据文件

将数据文件导入到 SPSS 中。

02 选择对话框

依次单击菜单“分析→均值比较→均值”，进入“均值”对话框。将变量“丝裂霉素浓度（ug/g）”指定为“因变量列表”；将变量“器官（organ）”、“时间（time）”和“药物（drug）”同时指定为“自变量列表”，如图 5.3 所示。

03 设置

在图 5.3 所示对话框中单击“选项”按钮，打开“均值分析：选项”对话框。在此对话框中，在“第一层的统计量”中选择“Anova 和 eta”选项，如图 5.4 所示。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
MEANS
   TABLES=cons  BY drug time organ
   /CELLS MEAN COUNT STDDEV
   /STATISTICS ANOVA .
```

05 单击“确定”按钮，指定均值分析的操作。

均值分析的结果解释如下：

- 基本描述统计量

表 5.3　描述性统计量

（a）药物分组的描述性统计量

丝裂霉素浓度（ug/g）

药　物	均　值	次　数	标准偏差
实验组	.916863	30	1.1745367
对照组	.338037	30	.1792969
总计	.627450	60	.8826428

（b）器官分组的描述性统计量

丝裂霉素浓度（ug/g）

器　官	均　值	次　数	标准偏差
心脏	.348770	30	.1624495
肝脏	.906130	30	1.1823323
总计	.627450	60	.8826428

（c）时间分组的描述性统计量

丝裂霉素浓度（ug/g）

时　间	均　值	次　数	标准偏差
15 分	.351015	20	.2711234
30 分	1.200495	20	1.3400775
60 分	.330840	20	.1765088
总计	.627450	60	.8826428

表 5.3 的（a）、（b）和（c）是未定义分组变量的描述性统计量，在“ 均值 ”主对话框中同时将 3 个分组变量选入“ 自变量列表 ”列表框中，因未定义分层变量，故按每个分类变量独立输出统计量。

表 5.4　分层后描述性统计量

丝裂霉素浓度（ug/g）

器　官	药　物	时　间	均　值	次　数	标准偏差
心脏	实验组	15 分	.114180	5	.0165175
		30 分	.332040	5	.0166926
		60 分	.253220	5	.0122017
		总计	.233147	15	.0942947
	对照组	15 分	.361320	5	.0110712
		30 分	.641200	5	.0245895
		60 分	.390660	5	.0078453
		总计	.464393	15	.1308657
	总数	15 分	.237750	10	.1309271
		30 分	.486620	10	.1641418
		60 分	.321940	10	.0730799
		总计	.348770	30	.1624495
肝脏	实验组	15 分	.777780	5	.0156473
		30 分	3.451560	5	.1175082
		60 分	.572400	5	.0100891
		总计	1.600580	15	1.3590444
	对照组	15 分	.150780	5	.0223150
		30 分	.377180	5	.0137081
		60 分	.107080	5	.0121197
		总计	.211680	15	.1235002
	总数	15 分	.464280	10	.3309571
		30 分	1.914370	10	1.6222589
		60 分	.339740	10	.2454740
		总计	.906130	30	1.1823323
总计	实验组	15 分	.445980	10	.3500767
		30 分	1.891800	10	1.6460343
		60 分	.412810	10	.1685534
		总计	.916863	30	1.1745367
	对照组	15 分	.256050	10	.1122001
		30 分	.509190	10	.1404108
		60 分	.248870	10	.1497694
		总计	.338037	30	.1792969

续表

器　官	药　物	时　间	均　值	次　数	标准偏差
	总计	15 分	.351015	20	.2711234
		30 分	1.200495	20	1.3400775
		60 分	.330840	20	.1765088
		总计	.627450	60	.8826428

表 5.4 为定义分层后描述性统计量，在“ 均值 ”对话框中将 3 个变量依次选入“ 自变量列表 ”列表框，每选入一个分组变量，便单击“ 下一个 ”按钮。当分组变量被定义为分层变量后，输出的统计量按选入分层变量的顺序将所有的因变量和分组变量完全列于一个表内，而不像未定义分层变量时那样每一个分组变量单列一个表。

- 方差分析表

表 5.5 为基于 3 个不同分组变量的各自独立的方差分析表。“ Between Groups ”表示变异来源的组间项，“ Within Groups ”表示变异来源的组内项，即误差项，“ Total ”表示总的变异来源。“ Sum of Squares ”为离均差平方和，“ Df ”为自由度，“ Mean Square ”为均方差，“ F ”表示 F 统计量，“ Sig. ”为 ρ 值，又称观察显著性水平。通过表中数据可以看出，两个药物组间的 ρ 值为 0.010，两个时间组间的 ρ 值为 0.001，两个器官组间的 ρ 值为 0.013，它们均小于 0.05，即两个药物组间、两个时间组间和两个器官组间都有显著性差异。

表 5.5　基于 3 个不同分组变量的各自独立的方差分析表

ANOVA Table

			Sum of Squares	Df	Mean Square	F	Sig.
丝裂霉素浓度（ug/g）*药物	Between Groups	(Combined)	5.026	1	5.026	7.120	.010
	Within Groups		40.939	58	.706		
	Total		45.964	59			

ANOVA Table

			Sum of Squares	Df	Mean Square	F	Sig.
丝裂霉素浓度（ug/g）*时间	Between Groups	(Combined)	9.855	2	4.928	7.779	.001
	Within Groups		36.109	57	.633		
	Total		45.964	59			

ANOVA Table

			Sum of Squares	Df	Mean Square	F	Sig.
丝裂霉素浓度（ug/g）*器官	Between Groups	(Combined)	4.660	1	4.660	6.543	.013
	Within Groups		41.305	58	.712		
	Total		45.964	59			

- 关联性统计量表

表 5.6　关联性统计量

Measures of Association

	Eta	Eta Squared
丝裂霉素浓度（ug/g）*药物	.331	.109

Measures of Association

	Eta	Eta Squared
丝裂霉素浓度（ug/g）*时间	.463	.214

Measures of Association

	Eta	Eta Squared
丝裂霉素浓度（ug/g）*器官	.318	.101

表 5.6 为关联性统计量表，用来描述药物、时间和器官的变化与丝裂霉素浓度变化的关联性，值越大，关联性也越大。从该表中可以看出，3 个因素的 Eta 统计量的 η^2 分别为 0.109、0.214 和 0.101。这说明在其他两个因素不变的情况下，时间的变化引起的丝裂霉素浓度变化最大。

5.2　方差分析概述

方差分析是 20 世纪 20 年代发展起来的一种统计方法，被广泛应用于分析心理学、生物学、工程和医药的试验数据等领域。在形式上方差分析是比较多个总体的均值是否相等，本质上它是研究变量之间的关系，这与回归分析方法有许多相同之处，但又有本质区别。在研究一个（或多个）分类性自变量与一个数值型因变量之间的关系时，方差分析是主要方法之一。本节主要介绍方差分析的基本方法和 SPSS 实现过程。

5.2.1　方差分析的基本原理

方差分析是研究分类性自变量对数值型因变量的影响，例如它们之间有无关系、关系强度如何等，方差分析方法首先检验总体的均值是否相等，进而判断分类型自变量对数值型因变量的显著性影响。

例如，在消费者市场，消费者和生产者之间常常因商品质量或服务质量发生纠纷。消费者在发生纠纷而无法解决时，便会向消费者协会投诉。为了评价不同行业的服务质量，消费者协会分别在零售业、旅游业、航空公司、家电制造业抽取了不同的企业作为样本，其中零售业 7 家，旅游业 6 家，航空公司 5 家，家电制造业 5 家，然后统计出近期消费者对这 23 家企业的投诉次数，表 5.7 给出了消费者对 4 个行业的投诉次数的统计数据列表。

表 5.7　消费者对 4 个行业的投诉次数

观 测 值	行　业			
	零 售 业	旅 游 业	航空公司	家电制造业
1	57	69	31	44
2	66	39	49	51
3	49	29	21	65
4	40	45	34	77
5	34	56	40	58
6	53	51		
7	44			

一般来说，企业收到的投诉次数越多，其服务质量也就越差。消费者协会想知道这几个行业之间的服务质量是否有显著差异，便对其进行了分析。

分析 4 个行业之间的服务质量是否有显著差异，即判断“ 行业 ”对“ 投诉次数 ”是否有显著影响，应该检验这 4 个行业被投诉次数的均值是否相等。如果均值相等，则认为“ 行业 ”对“ 投诉次数 ”是没有影响的，即它们之间的服务质量没有显著差异；反之，则认为行业之间的服务质量有显著差异。

上述问题可以通过方差分析得到解决。

方差分析是通过对数据误差来源的分析来判断不同总体之间的均值是否相等，进而分析自变量是否有影响。因此，进行方差分析时，区分误差的不同来源是重要的。下面给出误差来源分析及方差分析基本思想。

- 在同一行业下，样本各观测值是不同的。例如，所抽取的 7 家企业之间被投诉的次数是不同的。由于企业是随机抽取的，因此它们之间的差异可以看成是由随机因素的影响造成的，或者说是由于抽样的随机性造成的，这种误差称为随机误差。
- 在不同行业下，各观测值也是不同的。这种差异可能是由于抽样的随机性造成的，也可能是由于行业本身造成的，后者所形成的误差是由系统性因素造成的，称之为系统误差。
- 数据误差是用平方和表示的。衡量因素同一水平下样本数据的误差，称为组内误差（within groups），如上例中，零售业的 7 家企业被投诉的次数的误差。衡量不同水平下的样本误差称为组间误差（between groups），例如，零售业、旅游业、航空公司、家电制造业之间的被投诉次数的误差。组内误差只包含随机误差，而组间误差既包含随机误差，也包含系统误差。
- 由误差来源的分析得知，判断行业对投诉次数是否有影响，就是检验被投诉次数存在差异的原因。如果这种差异主要是系统误差，那么我们说不同行业对投诉次数有显著影响。在方差分析的前提下，这实际上就是检验 4 个行业被投诉次数的均值是否相等的问题。

- 根据统计学原理，组间均方和组内均方的比值构成 F 分布。给定显著性水平，通过和 F 分布统计量的概率 ρ 的比较，推断出总体均值是否存在显著性差异。

方差分析方法很多，不同方差构造的统计量也不同，本章后面几节将介绍单因素方差分析、多因素方差分析和协方差分析的基本原理和 SPSS 实现过程。

5.2.2　方差分析的概念和假设

在方差分析过程中，有几个重要的基本概念：

- 因素

在方差分析中，所要检验的对象称为因素。例如，在上例中“ 行业 ”是因素。

- 水平

因素的不同表现称为水平。上例中的零售业、旅游业、航空公司、家电制造业是行业因素的水平。

- 观测值

每个行业中得到的样本数据称为观测值。上例中只有“ 行业 ”一个因素，所以也称为单因素 4 行业因素水平的试验。

方差分析一般应满足三个基本假设，即要求各个总体应服从正态分布，各个总体的方差应相同以及观测值是独立的。

根据上述三个假设，方差分析对各个总体的显著性差异的推断就可转化为对各个总体均值是否存在差异的推断。

假设因素有 k 个水平，每个水平的均值分别用 $\mu_1,\mu_2,...\mu_k$ 表示，要检验 k 个水平的均值是否相等，需要提出如下假设检验问题：

$H_0:\mu_1=\mu_2=...=\mu_k$

$H_1:\mu_1,\mu_2,...\mu_k$ 不全相等。

例如，在上例中，我们假设 4 个行业被投诉次数均值分别为 μ_1，μ_2，μ_3，μ_4。为检验行业对投诉次数是否有影响，需要提出如下假设：

$H_0:\mu_1=\mu_2=\mu_3=\mu_4$

$H_1:\mu_1,\mu_2,\mu_3,\mu_4$ 不全相等。

与假设检验方法相比，方差分析不仅可以提高检验效率，同时由于它是将所有的样本信息都结合在一起，还增加了分析的可靠性。

5.3　单因素方差分析

SPSS 的“ 均值比较 ”模块提供了单因素方差分析的过程，下面介绍其基本操作和各个对话框的功能。

SPSS 单因素方差分析的具体步骤如下。

- 打开单因素方差分析对话框

依次单击菜单“ 分析→均值比较→单因素 ANOVA ”，进入“ 单因素方差分析 ”对话框，如图 5.5 所示。



图 5.5 “ 单因素方差分析 ” 对话框

- 选入变量

在变量列表框中选择因变量，单击 按钮，将其选入到右侧的“ 因变量列表 ” 列表框，可以选择一个以上的变量；在变量列表框中选择因子变量，单击 按钮，将其移动到“ 因子 ” 列表框。

- “ 对比 ” 选项

单击“ 对比 ” 选项，打开“ 单因素 ANOVA：对比 ” 对话框，如图 5.6 所示。

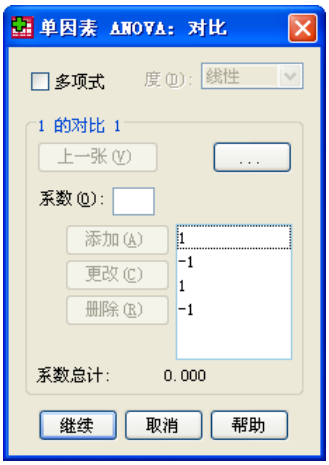


图 5.6 “ 单因素 ANOVA：对比 ” 对话框

“ 多项式 ” 选项用于将组间平方划分为趋势成分，可对因变量按因子变量中的水平次序进行趋势性检验。

“ 度 ” 下拉式菜单用于设定多项式的次数，可选项有：线性、二次项、立方、四次项和五次项。

“ 对比 ”选项区域中的“ 系数 ”需要用户指定检验对比将通过 T 检验，可以输入因变量中每组的系数，系数的顺序对应于因子变量中分类值的递增顺序，列表中的第一个系数对应于分类变量的最小值，最后一个系数对应于最大值。

- “ 两两比较 ”选项

单击“ 两两比较 ”选项，打开“ 单因素 ANOVA：两两比较 ”对话框，如图 5.7 所示。



图 5.7 “ 单因素 ANOVA：两两比较 ”对话框

在该对话框中进行均值的两两比较，各选项的含义如下。

“ 假定方差齐性 ”选项组给出当方差相等时确定两两比较的方法，包括 14 种可选方法，具体含义如表 5.8 所示。

表 5.8　假定方差相等各选项含义

方 法 名	含　义
LSD	Least-significant difference，即最小显著性差异法。用 T 检验完成各组间的配对比较，检验的敏感性高，各个水平间的均值存在微小的差异也有可能被检验出来，但此方法对第一类弃真错误的概率不进行控制和调整
Bonferroni	LSDMOD 方法，即修正最小显著差异法。用 T 检验完成各组间的配对比较；同时，通过设置每个检验的误差率来控制第一类错误的概率
Sidak	Sidak 法，根据 *t* 统计量进行配对多重比较，调整多重比较的显著性水平，其界限比 Bonferroni 方法小
Scheffe	Scheffe 差别检验法。α 可指定 0~1 之间任何显著性水平，默认值为 0.05
R-E-G-WF	Ryan-Einot-Gabriel-Welsch F 方法，根据 *F* 统计量检验的 Ryan-Einot-Welsch 多重比较
R-E-G-WQ	Ryan-Einot-Gabriel-Welsch Q 方法，根据 Student 极差统计量的 Ryan-Einot-Welsch 多重比较
S-N-K	Student-Newman-Keuls 方法，根据 Student 极差统计量在均值间进行配对比较；如果各组样本含量相等或者选择了“Harmonic average of all groups”(所有个组样本含量的调和平均值)，即用所有各组样本含量的调和平均值进行样本量的估计时，还用逐步过程进行齐次子集(差异较小的子集)的均值配对比较；在该比较过程中各组均值按从大到小的顺序排列，最先比较最末端的差异

续表

方 法 名	含　义
Tukey	Tukey's honestly significant difference 方法，即 Tukey 显著差异法，用 Student-Range 统计量进行所有组间均值的配对比较，用所有配对比较误差率作为实验误差率
Tukey's-b	Tukey-b 方法，用 Student-Range 分布进行组间均值的配对比较，其精确值为两种检验相应值的平均值
Duncan	Duncan 方法，指定一系列的 Range 值逐步进行计算比较得到结论
Hochbergs GT2	Hochberg's GT2 方法，用正态最大系数进行多重比较
Gabriel	Gabriel 方法，用正态标准系数进行配对比较，单元较大时这种方法较自由
Waller-Duncan	Waller-Duncan 方法，用 *t* 统计量进行多重比较检验，使用贝叶斯过程的多重比较检验
Dunnett	Dunnett 方法，指定此方法进行配对比较，方法是选择最后一组为对比其他各组和它比较，选定此方法后激活“ Control Catetory ”栏，在该栏展开下拉菜单选择对比组，其中可选“ Last ”分类和“ First ”分类；并在“ Test ”选项组中设定单双侧检验的选择，其中可选“ 2-Sided、<Control、>Control ”

“ 未假定方差齐性 ”选项组中给出当方差不齐时两两比较方法的选项，包括表 5.9 中列出的四项。

表 5.9　不假定方差相等各选项含义

方 法 名	含　义
Tamhane's T2	用 T 检验进行配对比较试验
Dunnett's T3	用基于 Student 最大模数的比较配对试验
Games-Howell	方差不齐时的配对比较试验
Dunnett's C	用基于 Student 极差的配对比较试验

- “ 选项 ”设置

单击“ 选项 ”按钮，打开“ 单因素 ANOVA：选项 ”对话框，如图 5.8 所示。



图 5.8 “ 单因素 ANOVA：选项 ”对话框

该对话框给出了检验统计量、检验和缺失值的处理方式等设置。

“ 统计量 ”选项组为输出统计量的选项，其选项及其含义如表 5.10 所示。

表 5.10　统计各选项含义

选 项 名	含　义
描述性	要求输出描述统计量，其中包括个案数目均值、标准差、标准误差、最小值、最大值以及各组中每个因变量的 95%置信区间
固定和随机效果	显示固定效应模型的标准差、标准误差和 95%的置信区间；随机效应模型的标准误差、95%的置信区间及方差成分间的估计值
方差同质性检验	要求计算分组树木相等的 Levene 统计量，该检验方法不依赖于正态假设
Brown-Forsythe	计算分组均数相等的 Brown-Forsythe 统计量。当不能把握方差齐性假设时，此统计量比 *F* 统计量更有优势
Welch	计算分组均数相等的 Welch 统计量。当不能把握方差齐性假设时，此统计量比 *F* 统计量更有优势

“ 均值图 ”对话框给出了均数图形选项“ 均值图 ”，选择该项，表示显示分组均数图。

“ 缺失值 ”选项组用于选择缺失值处理方法，如表 5.11 所示。

表 5.11　缺失值处理方法

处理方法	含　义
按分析顺序排除个案	剔除各分析中含有缺失值的个数
按列表排除个案	剔除含缺失值的全部个案

- 执行操作

在“ 单因素 ANOVA ”主对话框中，单击“ 确定 ”按钮，执行操作。

【示例 5.2】来自 4 个不同行业的投诉数据。

针对表 5.7 给出的数据进行单因素方差分析，并利用 LSD（最小显著差异检验）和 Bonferroni（修正的最小显著差异检验）方法，检验各行业间均值是否存在显著性差异；同时，选择线性多项式比较各均值，共指定两组多项式系数，检验行业对投诉次数的效应，多项式系数由下式给出。

$$1.0\times\text{mean}1-1.0\times\text{mean}2-1.0\times\text{mean}3+1.0\times\text{mean}4$$

用于检验零售业和家电业的投诉之和与旅游业和航空公司之间的投诉之和是否有显著差异。

$$1.0\times\text{mean}1+1.0\times\text{mean}2-1.0\times\text{mean}3-1.0\times\text{mean}4$$

用于检验零售业和旅游业的投诉之和与家电业和航空公司之间的投诉之和是否有显著差异。

（1）打开数据文件。

其中值 1、2、3、4 分别代表零售业、旅游业、航空业和家电业（如表 5.12 所示）。

（2）选择对话框

依次单击菜单“ 分析→均值比较→单因素 ANOVA ”，进入“ 单因素方差分析 ”对话框。变量“ 投诉次数 ”指定为“ 因变量列表 ”，变量“ 行业 ”指定为“ 因子 ”，如图 5.5 所示。

（3）设置

在图 5.5 所示对话框中单击“ 对比 ”按钮，打开“ 单因素 ANOVA：对比 ”对话框。在该对话框中选择“ 多项式 ”，激活“ 度 ”下拉菜单，默认选择“ 线性 ”，在“ 系数 ”文本框中依次输入多项式系数“ 1、-1、1、-1 ”和“ 1、1、-1、-1 ”，输入完毕后，单击“ 添加 ”按钮确认设置。如图 5.6 所示。

单击“ 两两比较 ”按钮，在“ 单因素 ANOVA：两两比较 ”对话框中依次选择“ LSD、Bonferroni、Sidak、Scheffe ”方法。

单击“ 选项 ”按钮，打开“ 单因素 ANOVA：选项 ”对话框。在此对话框中，依次选择“ 描述性 ”、“ 方差同质性检验 ”、“ 均值图 ”选项，如图 5.8 所示。

（4）单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
ONEWAY
     投诉次数 BY 行业
     /CONTRAST= 1 -1 1 -1
     /STATISTICS DESCRIPTIVES HOMOGENEITY
     /PLOT MEANS
     /MISSING ANALYSIS
     /POSTHOC = SCHEFFE LSD BONFERRONI SIDAK ALPHA(.05).
```

（5）单击“ 确定 ”按钮，指定单因素方差分析的操作。

单因素方差分析的结果解释如下：

- 基本描述性统计量

表 5.12　基本描述性统计量

Descriptives

投诉次数

	N	Mean	Std.Deviation	Std.Error	95%Confidence Interval for Mean		Minimun	Maximum
					Lower Bound	Upper Bound		
1	7	49.0000	10.80123	4.08248	39.0105	58.9895	34.0	66.0
2	6	48.0000	13.59412	5.54977	33.7338	62.2662	29.0	68.0
3	5	35.0000	10.41633	4.65833	22.0664	47.9336	21.0	49.0
4	5	59.0000	13.74755	5.70088	43.1718	74.8282	44.0	77.0
Total	23	47.0000	13.75865	2.86888	41.9199	53.8193	21.0	77.0

注释：表 5.12 中，N 为样本数量，Mean 为样本均值，Std.Deviation 为标准差偏离指数，Std.Error 为标准差误差，Confidence Interval for Mean 为置信区间（Lower Bound 表示置信区间的下限，Upper Bound 表示置信区间的上限），Minimum 和 Maximum 分别表示样本的最小值和最大值。

表 5.12 是基本描述性统计量。从表中可以看出，在 4 个行业中样本数量分别为 7、6、

5、5，其中家电业投诉次数最多，零售业和旅游业相近，航空业投诉最少，这一点也可以通过均值曲线图得到验证。

- 方差齐性检验

表 5.13　方差齐性检验结果

Test of Homogeneity of Variances

投诉次数

Levene Statistic	df1	df2	Sig.
.195	3	19	.898

表 5.13 是方差齐性的检验结果。从表中可以看出，在方差齐性检验计算出的 ρ 值为 0.898，在显著性水平为 0.05 的前提下，通过方差齐性检验，即不同行业投诉认为来自于相同方差的不同总体，以满足方差分析的前提。

- LSD 和 Bonferroni 检验后两两比较

表 5.14　两两比较检结果

Multiple Comparisons

Dependent Variable:投诉次数

			Mean Difference			95% Confidence Interval	
	(I) 行业	(J) 行业	(I-J)	Std. Error	Sig.	Lower Bound	Upper ound
Scheffe	1	2	1.00000	6.64193	.999	-19.3443	21.3443
		3	14.00000	6.99043	.292	-7.4118	35.4118
		4	-10.00000	6.99043	.574	-31.4118	11.4118
Scheffe	2	1	-1.00000	6.64193	.999	-21.3443	19.3443
		3	13.00000	7.22908	.382	-9.1428	35.1428
		4	-11.00000	7.22908	.524	-33.1428	11.1428
	3	1	-14.00000	6.99043	.292	-35.4118	7.4118
		2	-13.00000	7.22908	.382	-35.1428	9.1428
		4	-24.00000*	7.55053	.040	-47.1274	-.8726
	4	1	10.00000	6.99043	.574	-11.4118	31.4118
		2	11.00000	7.22908	.524	-11.1428	33.1428
		3	24.00000*	7.55053	.040	.8726	47.1274
LSD	1	2	1.00000	6.64193	.882	-12.9017	14.9017
		3	14.00000	6.99043	.060	-.6311	28.6311
		4	-10.00000	6.99043	.169	-24.6311	4.6311
	2	1	-1.00000	6.64193	.882	-14.9017	12.9017
		3	13.00000	7.22908	.088	-2.1306	28.1306
		4	-11.00000	7.22908	.145	-26.1306	4.1306

续表

	(I) 行业	(J) 行业	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper ound
	3	1	-14.00000	6.99043	.060	-28.6311	.6311
		2	-13.00000	7.22908	.88	-28.1306	2.1306
		4	-24.00000	7.55053	.005	-39.8034	-8.1966
	4	1	10.00000	6.99043	.169	-4.6311	24.6311
		2	11.00000	7.22908	.145	-4.1306	26.1306
		3	24.00000*	7.55053	.005	8.1966	39.8034
Bonferroni	1	2	1.00000	6.64193	1.000	-18.5532	20.5532
		3	14.00000	6.99043	.358	-6.5791	34.5791
		4	-10.00000	6.99043	1.000	-30.5791	10.5791
	2	1	-1.00000	6.64193	1.000	-20.5532	18.5532
		3	13.00000	7.22908	.528	-8.2817	34.2817
		4	-11.00000	7.22908	.867	-32.2817	10.2817
	3	1	-14.00000	6.99043	.358	-34.5791	6.5791
		2	-13.00000	7.22908	.528	-34.2817	8.2817
		4	-24.00000*	7.55053	.030	-46.2280	-1.7720
	4	1	10.00000	6.99043	1.000	-10.5791	30.5791
		2	11.00000	7.22908	.867	-10.2817	32.2817
		3	24.00000*	7.55053	.030	1.7720	46.2280
Sidak	1	2	1.00000	6.64193	1.000	-18.4891	20.4891
		3	14.00000	6.99043	.309	-6.5117	34.5117
		4	-10.00000	6.99043	.670	-30.5117	10.5117
Sidak	2	1	-1.00000	6.64193	1.00	-20.4891	18.4891
		3	13.00000	7.22908	.425	-8.2119	34.2119
		4	-11.00000	7.22908	.608	-32.2119	10.2119
	3	1	-14.00000	6.99043	.309	-34.5117	6.5117
		2	-13.00000	7.22908	.425	-34.2119	8.2119
		4	-24.00000*	7.55053	.029	-46.1551	-1.8449
	4	1	10.00000	6.99043	.670	-10.5117	30.5117
		2	11.00000	7.22908	.608	-10.2119	32.2119
		3	24.00000*	7.55053	.029	1.8449	46.1551

*The mean difference is significant at the .05 level.

表 5.14 是利用 LSD、Bonferroni、Sidak 和 Scheffe 检验方法分别显示两两行业之间投诉次数均值的检验比较结果。表中的星号表示在显著性水平为 0.05 的条件下，相应的两组均值存在显著差异。

各种检验方法对抽样分布标准误差的定义不尽相同，但在系统中旨在采用 LSD 方法的标准误差，故表中两种方法的两列数据完全相同，第三列和 Sig 是检验统计量的观测值在不同分布中的概率 ρ 值。

两种方法存在一定的差异，两者之间由于对误差率的控制不同，所以敏感度也不同，从表中可以明显地看出，LSD 方法的概率 ρ 值都比 Bonferroni 方法的相应概率 ρ 值小一些，与其他方法相比，LSD 方法的敏感度是比较高的。

例如，在显著性水平为 0.05 的前提下，LSD 检验中航空业与家电业之间的投诉次数均值存在显著差异，其概率 ρ 值为 0.005，Bonferroni 方法中两者之间虽然也存在显著性差异，但其统计量的概率 ρ 值为 0.030，远远大于 LSD 方法的概率 ρ 值。

- 多项式检验结果

表 5.15 给出了线性多项式比较系数，表 5.16 给出了比较检验的结果。

表 5.15　线性多项式比较系数

Contrast Coefficients

Contrast	行　业			
	1	2	3	4
1	1	-1	1	-1
2	1	1	-1	-1

由计算得到的概率值可知，在 Contrast1 的情形下，无论假设为方差齐性，还是方差不齐，都有 $\rho < 0.05$，小于显著性水平，拒绝零假设，即认为零售业、家电业投诉次数之和与旅游业、航空业投诉次数之和之间在 5%的水平上差异显著；在 Contrast2 的情形下，$\rho > 0.05$，大于显著性水平，接受零假设，即认为零售业、旅游业投诉次数之和与家电业、航空业投诉次数之和之间差异显著。

表 5.16　多项式系数比较检验结果

Contrast Tests

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
投诉次数	Assume equal variances	1	-23.0000	10.05613	-2.287	19	.034
		2	3.0000	10.05613	.298	19	.769
	Does not assume equal variances	1	-23.0000	10.08299	-2.281	16.730	.036
		2	3.0000	10.08299	.298	16.730	.770

实际上，从两个行业均值之差可以明显得出此结论，后者均值之差为 3，但前者却达到了 25，有较明显的差异。

- 单因素方差分析表

表 5.17　单因素方差分析表

ANOVA

投诉次数

		Sum of Squares	df	Mean Square	F	Sig
Between Groups	(Combined)	1456.609	3	485.536	3.407	.039
	Liner Term Unweighted	83.710	1	83.710	.587	.453
	Weighted	52.174	1	52.174	.366	.552
	Deviation	1404.435	2	702.217	4.927	.019
Within Groups		2708.000	19	142.526		
Total		4164.609	22			

注释：表 5.17 中，Sum of Square 表示偏差平方和，df 为样本自由度，Mean Square 为偏差均方值，F 为统计量 F 值，Sig.为自由度的熵。Between Groups 和 Within Group 分别表示方差的两个来源——组内和组间。

表 5.17 是单因素方差分析表。从表中可以看出，第五列是统计量 F 值，计算得 3.407；表中第二行和第三行给出方差的两个来源，即“ Between Groups ”和“ Within Group ”；第三列是检验统计量的自由度（df），组间自由度为 3，组内自由度为 19；总自由度为为 22。

第二列表示偏差平方和，其中组间偏差平方和为 14556.609，组内偏差平方和为 2708.000，总偏差平方和为 4164.609。

均方为偏差平方和与自由度的熵，分别为 485.536 和 142.526，两者之比为 F 分布的观测值 3.407。针对假设 H_0，组间均值存在显著性差异，计算 F 分布的观测值 3.407，而对应的概率 ρ 值为 0.039。在默认显著性水平为 0.05 的前提下，由于 F 分布的观测值大于概率 ρ 值，故应拒绝原假设，即认为不同行业之间的投诉次数差异显著。

- 两两比较的相似形子集

表 5.18　两两比较的相似形子集

投诉次数

行　业		N	Subset for alpha=.05	
			1	2
Scheffe[a,b]	3	5	35.0000	
	2	6	48.0000	48.0000
	1	7	49.0000	49.0000
	4	5		59.0000
	Sig.		.306	.510

Means for groups in homogeneous subsets are displayed.

a．Uses Harmonic Mean Sample Size = 5.638.

b．The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

注释：表 5.17 是由 Scheffe 方法划分的相似性子集。从表中可以看出，在显著性水平为 0.05

（alpha=0.05）的情况下，第一组包括的行业有 3（航空公司）、2（旅游业）、1（零售业）；而第二组包括的行业有 2（旅游业）、1（零售业）和 4（家电制造业）；其中旅游业、零售业为两个组共有，其均值在两组中差距近似，但航空公司和家电制造业的均值差较大。第一个子集中，组内相似的概率为 0.306（Sig.=0.306），第二组的组内相似性概率为 0.510（Sig.=0.510）。

- 均值连线图

图 5.9 是选择“ 均值连线 ”选项后的均值连线图。图中各点表示 4 个行业对应数据的均值大小。

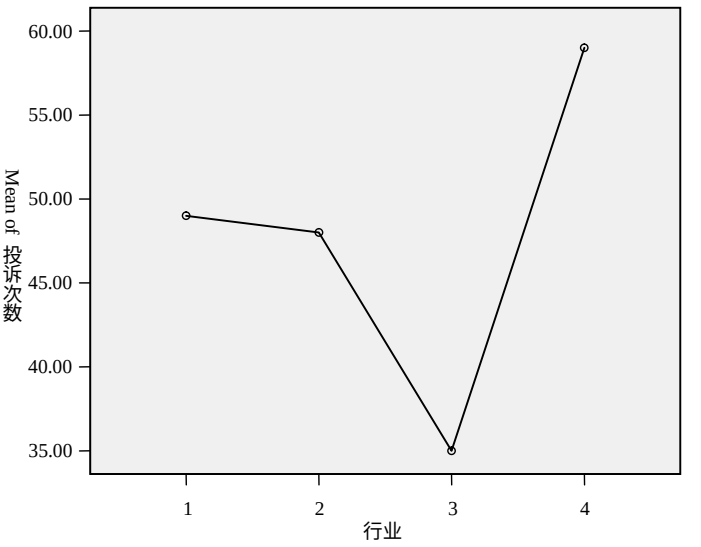


图 5.9　均值连线图

5.4　多因素方差分析

多因素方差分析用来研究两个及两个以上的控制变量是否对观察变量产生显著影响。由于讨论多个因素对个案的影响，因此这种方差分析过程被称为多因素方差分析。多因素方差分析不仅能够分析多个因素对观测变量的影响，而且能够分析多个控制因素的交互作用能否对观测变量的分布产生影响，进而最终找到利于观测变量的最优组合。

例如，医学研究中需要对若干个男人和若干个女人头部的 4 个典型解剖部位进行测量，研究男、女头部有无显著差异；又如在冶金行业中研究片状钛合金在不同温度下、不同拉伸速度下的抗拉性能是否一致。这些问题都属于多因素方差分析。

多因素方差分析的基本原理与单因素方法分析基本相似，都是通过检验两个或多个样本均值间是否存在显著差异而得出结论的。

SPSS 利用“ 一般线性模型 ”模块的“ 单变量 ”过程来完成多因素方差分析。“ 单变量 ”过程可以处理各种不同设计，如完全随机区组设计、析因设计、拉丁方设计、裂区试验设

计、交叉设计和协方差分析。

SPSS 多因素方差分析的具体步骤如下。

● 打开多因素方差分析对话框

依次单击菜单“分析→一般线性模型→单变量”，进入“单变量”对话框，如图 5.10 所示。



图 5.10 “单变量”对话框

● 选入变量

在变量列表框中选择因变量，单击 按钮，将其选入到右侧的“因变量”列表框；在变量列表框中选择固定因素变量，单击 按钮，将其移动到“固定因子”列表框；如果需要，选择随机因素变量移动到“随机因子”列表框；选择协变量移动到“协变量”列表框。

● “模型”选项

在图 5.10 所示对话框中单击“模型”按钮，打开“单变量：模型”对话框，如图 5.11 所示。

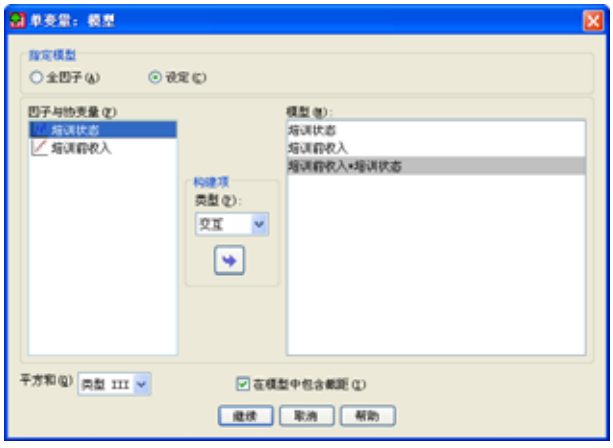


图 5.11 “单变量：模型”对话框

“ 指定模型 ” 选项组指定模型的类型，包括表 5.19 中的两个选项。

表 5.19　指定模型各选项含义

选　　项	含　　义
全因子	默认选项。选择该项，表示建立全因素模型。全因素模型包括所有因素主效应、所有协变量主效应及所有因素与因素的交互效应，不包括协变量的交互效应
设定	自定义的模型。只指定一个交互效应的子集或因素与协变量的交互效应。用户必须指定模型的所有项目

选择了 “ 设定 ” 选项以后，在 “ 因子与协变量 ” 列表框中自动列出可以作为因素变量的变量名，名称后用括号内的字母 F 标识；也列出可以作为协变量的变量名，名称后用括号内的字母 C 标识。

单击 “ 构建项 ” 中的下拉箭头，在出现的下拉菜单中选择要构建的类型，其各选项的含义如表 5.20 所示。

表 5.20　构建项的各选项含义

选　　项	含　　义
交互效应	建立所有被选变量最高水平的交互效应项目，默认选项
主效应	建立每个被选变量主效应
所有二阶	建立所有 2 阶交互效应
所有三阶	建立所有 3 阶交互效应
所有四阶	建立所有 4 阶交互效应
所有五阶	建立所有 5 阶交互效应

该对话框 “ 平方和 ” 下拉列表用于选择平方和的分解方法，单击下拉箭头，出现下拉菜单，包括如表 5.21 所示选项。

表 5.21　“ 平方和 ” 下拉列表的各选项含义

选　　项	含　　义
类型 I	Ⅰ型平方和，适用于平衡方差分析模型、多项式回归模型、完全嵌套设计
类型 II	Ⅱ型平方和，适用于平衡方差分析模型、只有一个主因子的模型、回归模型、完全嵌套设计
类型III	Ⅲ型平方和，适用于Ⅰ型、Ⅱ型平方和适用的所有模型、无缺失值的所有平衡与不平衡模型。此项较常用，为默认选项
类型IV	Ⅳ型平方和，适用于平Ⅰ型、Ⅱ型平方和适用的所有模型、无缺失值的所有平衡与不平衡模型

“ 在模型中包含截距 ” 选项表示的含义是：通常截距包括在模型汇总中，如果能假设数据通过原点，则可以不包括截距，即不选择此项。

- “ 对比 ” 选项

在图 5.10 所示对话框中单击 “ 对比 ” 按钮，打开 “ 单变量：对比 ” 对话框，如图 5.12 所示。



图 5.12 “ 单变量：对比 ” 对话框

在“ 因子 ” 列表框中显示出所有在主对话框中选中的因素变量，因素变量名后的括号是当前的对比方法。可以在“ 更改对比 ” 选项区域中改变对比方法，方法是在“ 因子 ” 列表框中选择需要改变对比方法的因子，然后单击“ 对比 ” 下拉箭头，在下拉菜单中选择如表 5.22 所示的对比方法：

表 5.22　各对比方法的含义

方　法	含　义
无	不进行均值比较
偏差	差别对比，比较每一水平的平均效应和总平均效应选择
简单	简单对比，每个水平与指定水平的平均效应比较，可选择第一组和最后一组作为参考水平
差值	差分对比，每个水平与前面各水平的平均效应的比较
Helmert	Helmert 对比，每个水平与随后的各水平的平均效应的比较
重复	重复对比，每个水平与随后的一个水平的比较
多项式	多项式对比，比较线行效应、二次效应、三次效应等，常用于预测多项式趋势

改变对比的设置完成后，单击“ 更改 ” 按钮，改变了的对比方法就显示在“ 因子 ” 列表框中。设置完毕后，单击“ 继续 ” 按钮，返回到主对话框。

- “ 绘图 ” 选项

在图 5.10 所示对话框中单击“ 绘制 ” 按钮，打开“ 单变量：轮廓图 ” 对话框，如图 5.13 所示。



图 5.13 “ 单变量：轮廓图 ” 对话框

在该对话框中可以绘制一个或多个因素变量为参考的因变量均值分布图。在只有一个因素水平时，为因变量估计边缘均值的轮廓图；在有两个以上因素水平时，绘制分离线。

“ 因子 ” 列表框显示因素的变量名，将该列表框中的因素变量移动到右侧不同列表框中，可选列表框有：水平轴、单图和多图。

“ 因子 ” 列表框的变量移动到相应的坐标轴窗口后，单击 “ 添加 ” 按钮，所选因素变量就出现在 “ 图 ” 列表框中。

- “ 保存 ” 选项

在图 5.10 所示对话框中单击 “ 保存 ” 按钮，打开 “ 单变量：保存 ” 对话框，如图 5.14 所示。

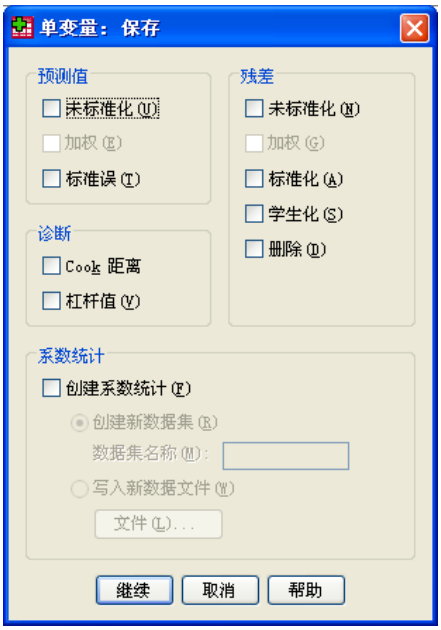


图 5.14 “ 单变量：保存 ” 对话框

“ 单变量：保存 ” 对话框给出预测值残差和检测值作为新的变量的保存选项，预测值各选项的含义如表 5.23 所示。

表 5.23　预测值各选项含义

选　项	含　义
未标准化	非标准化预测值
加权	加权非标准化预测值。如果在主对话框中选择 WLS（加权最小二乘）变量，保存因变量的加权非标准化预测值
标准误差	预测值均值的标准误差

“ 诊断 ” 选项组给出诊断结果保存选项，各选项的含义如表 5.24 所示。

表 5.24　诊断各选项含义

选　项	含　义
Cook 距离	Cook 距离
杠杆	非中心化杠杆值

“ 残差 ”选项组给出诊断结果保存选项，各选项的含义如表 5.25 所示。

表 5.25　残差各选项含义

选　项	含　义
未标准化	非标准化残差值。残差为观测值与预测值之差
加权	权重非标准化残差。如果在主对话框中选择 WLS 变量，则保存权重非标准化残差
标准化	标准化残差，即 pearson 残差
学生化	学生化残差
删除	删除残差自变量值与调整预测值之差

- “ 选项 ”设置

在图 5.10 所示对话框中单击“ 选项 ”按钮，打开“ 单变量：选项 ”对话框，如图 5.15 所示。



图 5.15 “ 单变量：选项 ”对话框

“ 估计边际平均值 ”选项组用于选择显示结果的因素变量，方法是在左侧的“ 因子与因子交互 ”列表框中选择因素和交互项，单击将其移动到“ 显示均值 ”列表框。

如果在“ 显示均值 ”栏中有主效应变量被选中，则可以选择“ 比较主效应 ”选项，对主效应变量进行估计边际均值，并在“ 置信区间调节 ”列表框中选择 LSD、Bonferroni 或

Sidar 方法。

“ 输出 ” 选项组指定要求输出的统计量，其选项含义如表 5.26 所示。

表 5.26　输出各选项含义

选　项	含　义
描述统计	输出描述统计量。显示所有单元格中因变量的观测均值、标准差和例数
功效估计	效应大小的估计。计算每个效应的估计及参数估计、标准误差、置信区间
检验效能	给出计算功效的显著性水平 Alpha 值，该值在 0.01~0.99 之间。显示观测功效系统，默认显著性水平是 0.05
参数估计	给出各因变量与自变量的回归
对比系数矩阵	将系数矩阵进行对比
方差齐性检验	方差齐次检验
分布-水平图	可绘制观测值与残差的散点图
残差图	绘制残差图
缺乏拟合优度检验	拟合度不足的检验。检查独立变量和非独立变量间的关系是否被充分描述
一般估计函数	可以根据一般估计函数自定义假设检验对比系数矩阵的行与一般函数的线性组合

在 “ 显著性水平 ” 文本框中输入显著性水平，用于检验两两比较的置信水平和计算检验的观察效能。

【示例 5.3】机器和工人两因素的方差分析。

3 台机器（利用 A、B、C 表示）生产同一种产品，表 5.27 给出了 4 名工人操作机器 A、B、C 生产的产品的产量，问机器之间、工人之间在产量上是否存在显著差异。

表 5.27　不同机器不同工人的产量数据

机器类型	1	2	3	4
A	50	47	47	53
B	63	54	57	58
C	52	42	41	48

利用 SPSS 进行多因素方差分析的基本步骤如下。

01 打开数据文件

变量 *x* 表示产量，变量 *a*、*b* 分别表示不同机器和不同工人，如图 5.16 所示。

02 选入变量

依次单击菜单 “ 分析→一般线性模型→单变量 ”，打开 “ 单变量 ” 对话框。将 “ 产量[x] ” 选入 “ 因变量 ” 列表框；将 “ 机器[a] ” 和 “ 工人[b] ” 选入 “ 固定因子 ” 列表框，如图 5.17 所示。

03 设置

单击 “ 模型 ” 按钮，在 “ 单变量：模型 ” 对话框中选择 “ 自定义 ” 选项；在 “ 建立条件 ” 下拉菜单中选择 “ 主效应 ”，然后在 “ 因子与协变量 ” 中将 “ a ” 和 “ b ” 引入 “ 模型 ” 列表框，如图 5.18 所示。

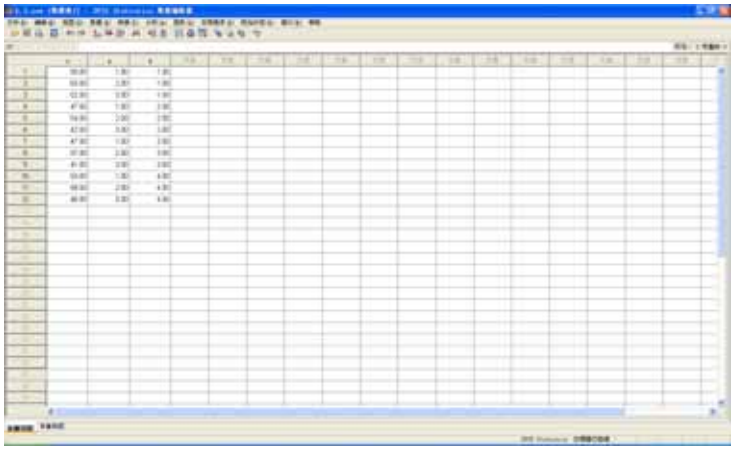


图 5.16　打开数据文件



图 5.17 “ 单变量 ” 对话框



图 5.18 “ 单变量：模型 ” 对话框

单击 “ 两两比较 ” 按钮，在 “ 单变量：观测均值的两两比较 ” 对话框的 “ 因子 ” 列表

框中选择变量 a 和 b，移动到“ 两两比较检验 ”列表框，并且在“ 假定方差齐性 ”栏中选中“ S-N-K ”选项。如图 5.19 所示。



图 5.19 “ 单变量：观测均值的两两比较 ”对话框

04 单击“ 粘贴 ”按钮，得到 Syntax 命令语句程序如下：

```
UNIANOVA
      x  BY a b
      /METHOD = SSTYPE(3)
      /INTERCEPT = INCLUDE
      /POSTHOC = a b ( SNK )
      /CRITERIA = ALPHA(.05)
      /DESIGN = a b .
```

05 单击“ 确定 ”按钮，执行单变量多因素方差分析的操作。

多因素方差分析的结果解释如下：

- 组间因素

表 5.28 给出各因素的水平数。分组变量 a 有 3 个水平，即机器 A、B、C；每个水平有 4 例；分组变量 b 有 4 个水平，即工人 1、2、3、4，每个水平有 3 例。

表 5.28　组间因素

Between-Subjects Factors

		N
机器	1．00	4
	2．00	4
	3．00	4
工人	1．00	3
	2．00	3
	3．00	3
	4．00	3

● 组间效应值

表 5.29 给出组间效应检验结果。从表中可以看出，因素 a 的检验统计量 F 的观测值为 29.102，检验的概率 ρ 值为 0.001，小于 0.05，拒绝零假设，可以认为机器之间存在差异显著，各机器之间的产量不全相等；因素 b 的检验统计量 F 的观测值为 6.985，检验的概率 ρ 值为 0.022，小于 0.05，拒绝零假设，可以认为工人之间存在差异显著，各工人之间的产量不全相等。

表 5.29　组间效应值

Tests of Between-Subjects Effects

Dependent Variable：产量

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	433.167[a]	5	86.633	15.831	.002
Intercept	31212.000	1	31212.000	5703.716	.000
a	318.500	2	159.250	29.102	.001
b	114.667	3	38.222	6.985	.022
Error	32.833	6	5.472		
Total	31678.000	12			
Corrected Total	466.000	11			

a．R Squared = .930 (Adjusted R Squared = .871)

● 因素 a 的多重验后检验

表 5.30 为因素 a 的多重验后检验结果。由于本例是无重复设计，不能求出每个格子的方差，故不能计算方差的齐次性。这里仅选择 S-N-K 法进行均值之间的两两比较。

表 5.30　因素 a 的多重验后检验产量

Srudent-Newman-Keuls[a,b]

机　器	N	Subset	
		1	2
3.00	4	45.7500	
1.00	4	49.2500	
2.00	1		58.0000
Sig.		.079	1.000

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 5.472.

a．Uses Harmonic Mean Sample Size = 4.000.

b．Alpha = .05.

注释：表 5.30 中，Sig.为自由度的熵，N 为样本数量，Subset 为均衡子集。

在均衡子集表中，第一均衡子集（Subset=1 列）包含第三组（a=3）和第一组（a=1），

它们的均值分别为 45.7500 和 49.2500，两组均值比较的概率 ρ 值为 0.079，大于 0.05，接受零假设，即可认为机器 C 和 A 的产量的均值之间无明显差异，而机器 B 和它们的差异较显著。

- 因素 b 的多重验后检验

表 5.31 为因素 b 的多重验后检验结果。从表中可以看出，第一均衡子集（Subset=1 列）包含第二组（a=2）、第三组（a=3）和第四组（a=4），它们的均值分别为 45.6667、48.3333 和 53.0000，两组均值比较的概率 ρ 值为 0.070，大于 0.05，接受零假设，即可认为工人 2、工人 3 和工人 4 的产量均值之间无明显差异。

表 5.31　因素 b 的多重验后检验产量

Student-Newman-Keuls

机　　器	N	Subset	
		1	2
2.00	3	47.6667	
3.00	3	48.3333	
4.00	3	53.0000	53.0000
1.00	3		55.0000
Sig.		.070	.335

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 5.472.

a．Uses Harmonic Mean Sample Size = 3.000.

b．Alpha = .05

第二均衡子集（Subset=2 列）包含第四组（b=4）和第一组（b=1），它们的均值分别为 53.0000 和 55.0000，两组均值比较的概率 ρ 值为 0.335，大于 0.05，接受零假设，即可认为工人 4 和工人 1 的产量均值之间无明显差异。

第一组和第二、三、四（b=2、3、4）组没有列在均衡子集表的同一个格子中，可以认为它们的均值存在显著差异。

5.5　重复度量

重复测量资料是指对同一受试对象的某项观测指标进行多次测量所得到的数据。如对病人治疗（后手术）后一天、三天、一周、两周等多个时间点进行连续观测；教育研究中观察不同学期学生的成绩变化情况；心理研究中观察不同时间段个体的心理调适能力；经济领域中研究市场的动态等。重复测量设计的数据分析若采用前述的普通方差分析方法，同样需要满足独立、正态、等方差的前提假设。可实际情况是：重复测量观测值来自同一受试对象的不同时间点，不能完全满足以上各项前提假设条件。

独立性：由于数据间相关性的存在，违背了方差分析要求数据满足“ 独立性 ”的基本

条件。在这种情况下，若使用一般的方差分析方法，将会增加犯 I 类错误的概率。

等方差性：对于重复测量数据，另外一个前提假设为方差齐性，要求各时间点测量值的方差相等，即独立结构相关系数相对应的协方差矩阵为球对称（Sphericity）结构。在生物、社会，尤其是行为、心理领域较少有满足球形条件的重复测量数据。这种前提条件的破坏直接影响到分析结果；但幸运的是，目前可以采用调整自由度的方法或多变量分析的方法来解决方差不等问题。球对称假设的检验可以采用“ Mauchly 检验、Box 检验、Greenhouse-Geisser 检验和 Huynh-Feldt 检验 ”。

正态性：要求重复测量资料必须服从正态分布，数据是否服从正态分布，可以依据经验做出判断；在样本含量不太小的情况下，方差分析对即使略偏离正态分布的资料，结果也较稳健。

实际上，重复测量设计并非单纯的一种设计方法，重复测量可以出现在实验设计、临床试验设计及调查设计中，但切不可以用传统的分析方法，比如，最简单的设计类型类似于随机区组设计，不可采用随机区组设计的方差分析、重复测量数据的方差理论会让非统计专业的科研工作者望而却步，但采用 SPSS 软件进行分析却十分方便。

【示例 5.4】分层随机抽样重复测量数据。

经营快餐的一家连锁店计划改进某一营业品种，提出了 3 种方案，并随机选择了若干个市场，每个市场有多个网点。要求在每个市场只能销售其中一种新品种，之后观察记录每个网点每周的销售量，连续观察 4 周。其他因素还包括市场规模、营业年限。结果数据如表 5.32 所示。

表 5.32　重复测量数据

市场编号	规　模	网　点	年　限	方　案	周　次	销 售 量
1	3	1	7	3	1	70.63
1	3	1	7	3	2	56.28
1	3	1	7	3	3	70.98
1	3	1	7	3	4	69.91
1	3	2	11	2	1	68.42
1	3	2	11	2	2	56.74
1	3	2	11	2	3	60.04
1	3	2	11	2	4	63.64
1	3	3	1	2	1	68.25
1	3	3	1	2	2	62.20
1	3	3	1	2	3	58.81
1	3	3	1	2	4	58.63
1	3	4	6	2	1	59.18
1	3	4	6	2	2	62.41
1	3	4	6	2	3	62.04
1	3	4	6	2	4	67.58
1	3	5	10	1	1	74.97
1	3	5	10	1	2	69.20

续表

市场编号	规　模	网　点	年　限	方　案	周　次	销 售 量
1	3	5	10	1	3	65.67
1	3	5	10	1	4	66.64
…	…	…	…	…	…	…
10	1	901	5	1	1	54.71
10	1	901	5	1	2	54.02
10	1	901	5	1	3	53.33
10	1	901	5	1	4	57.65
10	1	902	4	3	1	34.63
10	1	902	4	3	2	47.06
10	1	902	4	3	3	48.42
10	1	902	4	3	4	49.28
10	1	903	11	3	1	48.33
10	1	903	11	3	2	50.50
10	1	903	11	3	3	49.48
10	1	903	11	3	4	44.35
10	1	904	13	1	1	45.09
10	1	904	13	1	2	54.07
10	1	904	13	1	3	58.04

5.5.1　数据重构

表 5.32 提供的数据是数据库的通用格式，对不熟悉数据库的读者来说，可能从中体会不到重复测量。另外，在进行重复测量数据分析时，需要对数据进行重新整理，即数据重构，将数据转换为重复测量数据所要求的格式。

依次单击菜单“ 数据→重组 ”，打开“ 重组数据向导 ”对话框，如图 5.20 所示。

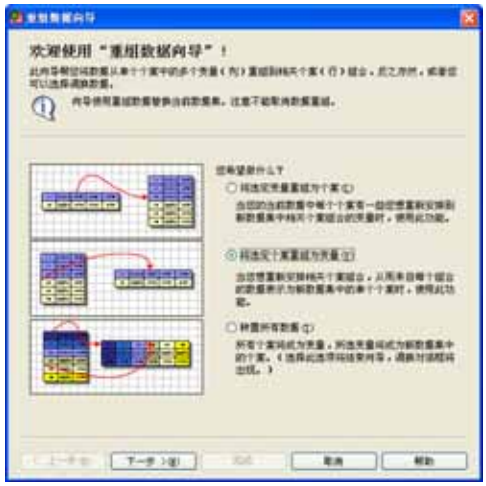


图 5.20 “ 重组数据向导 ”对话框

选择第二项“将选定个案重组为变量”，单击“下一步”按钮，得到如图 5.21 所示的“重组数据向导-第 2 步（共 5 步）”对话框。定义网点变量为数据库标识变量，定义周次为重复测量标识变量。

依次单击“下一步→完成”按钮。

重构后的数据库如表 5.33 所示。

表 5.33　重构后的数据库

市场编号	规 模	网 点	年 限	方 案	销 售 量			
					第 1 周	第 2 周	第 3 周	第 4 周
1	1	1	7	3	70.63	56.28	70.98	69.91
1	1	2	11	2	68.42	56.74	60.04	63.64
1	1	3	1	2	68.25	62.20	58.81	58.63
1	1	4	6	2	59.18	62.41	62.04	67.58
1	1	5	10	1	74.97	69.20	65.67	66.64
…	…	…	…	…	…	…	…	…
10	1	903	11	3	48.33	50.50	49.48	44.35
10	1	904	13	1	45.09	54.07	58.04	46.82



图 5.21 “重组数据向导-第 2 步（共 5 步）”对话框

表 5.33 为数据重构后的数据。比较表 5.32 和表 5.33 可见，共有 904 个网点分布在 10 个市场，每个网点连续 4 周记录销售量，所以表 5.32 共有 904×4=3616 个销售量数据值。如果将每一个网点 4 周记录销售量放在一行，则表 5.32 便变成了表 5.33。

不同市场的快餐连锁店某品种 3 种改进方案下的营业网点个数的分布见表 5.34。由此

可见，不同市场、不同方案下的网点数不等，所以该数据为不平衡数据。

表 5.34　不同方案的每个市场网点数分布

市场编号	方　案			合　计
	1	2	3	
1	4	7	10	21
2	6	4	8	18
3	6	4	5	15
4	2	1	2	5
5	8	8	7	23
6	1	3	0	4
7	4	4	1	9
8	2	5	4	11
9	6	11	6	23
10	2	0	2	4
合计	41	47	45	133

5.5.2　重复度量

数据重构完成以后，就可以进行重复度量了，具体步骤如下。

01 选择重复测量方差对话框

依次单击菜单“ 分析→一般线性模型→重复度量 ”命令。

02 选入变量

在“ 被试内因子名称 ”文本框中输入“ week ”，在“ 级别数 ”文本框中输入“ 4 ”，单击“ 添加 ”按钮，将其添加到右边的列表框中；在“ 度量名称 ”文本框中输入“ sales ”，单击“ 添加 ”按钮，将其加入到右边的列表框中，如图 5.22 所示。



图 5.22 “ 重复度量定义因子 ”对话框

03 设置

单击图 5.22 对话框上“ 定义 ”按钮，打开“ 重复度量 ”对话框。将“ sales.1、sales.2、sales.3、sales.4 ”选入“ 群体内部变量 ”列表框；将“ MaketID、Promotion ”选入“ 因子列表 ”列表框，如图 5.23 所示。



图 5.23 “ 重复度量 ”对话框

单击如图 5.23“ 重复度量 ”对话框上的“ 模型 ”按钮，打开“ 重复度量：模型 ”对话框。“ 指定模型 ”选择“ 全因子 ”选项；“ 平方和 ”选项选择“ 类型Ⅳ ”，如图 5.24 所示。



图 5.24 “ 重复度量：模型 ”对话框

单击如图 5.23 所示的“ 重复度量 ”对话框上的“ 绘制 ”按钮，打开“ 重复度量：轮廓图 ”对话框。将“ week ”选入“ 水平轴 ”列表框；将“ promo ”选入“ 单图 ”列表框，如图 5.25 所示。



图 5.25 “ 重复度量：轮廓图 ”对话框

单击如图 5.23 所示的“ 重复度量 ”对话框上的“ 选项 ”按钮，打开“ 多变量：选项 ”对话框。选择“ 功效估计、SSCP 矩阵和方差齐性检验 ”选项。如图 5.26 所示。

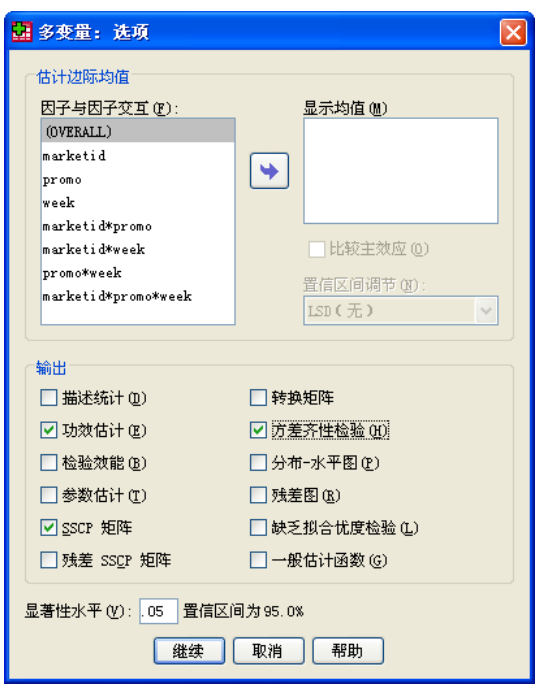


图 5.26 “ 多变量：选项 ”对话框

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
GLM
   sales.1 sales.2 sales.3 sales.4 BY marketid promo
   /WSFACTOR = week 4 Polynomial
   /MEASURE = sales
   /METHOD = SSTYPE(4)
   /PRINT = ETASQ TEST(SSCP) HOMOGENEITY
   /CRITERIA = ALPHA(.05)
   /WSDESIGN = week
   /DESIGN = marketid promo marketid*promo .
```

05 单击“确定”按钮，执行重复度量的操作。

重复度量的结果解释如下：

- SSCP 矩阵分析

表 5.35　SSCP 矩阵

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
week	Pillai's Trace	.056	2.033[a]	3.000	103.000	.114	.056
	Wilks' Lambda	.944	2.033[a]	3.000	103.000	.114	.056
	Hotelling's Trace	.059	2.033[a]	3.000	103.000	.114	.056
	Roy's Largest Root	.059	2.033[a]	3.000	103.000	.114	.056
week	Pillai's Trace	.210[b]	.877	27.000	315.000	.646	.070
	Wilks' Lambda	.803[b]	.870	27.000	301.455	.656	.070
	Hotelling's Trace	.229[b]	.863	27.000	305.000	.666	.071
	Roy's Largest Root	.109[b]	1.270[c]	9.000	105.000	.262	.098
week * promo	Pillai's Trace	.038[b]	.678	6.000	208.000	.668	.019
	Wilks' Lambda	.962[b]	.672[a]	6.000	206.000	.672	.019
	Hotelling's Trace	.039[b]	.666	6.000	204.000	.677	.019
	Roy's Largest Root	.026[b]	.885[c]	3.000	104.000	.452	.025
week * marketid * promo	Pillai's Trace	.409	1.037	48.000	315.000	.413	.136
	Wilks' Lambda	.638	1.043	48.000	307.142	.403	.139
	Hotelling's Trace	.495	1.049	48.000	305.000	.393	.142
	Roy's Largest Root	.295	1.935[c]	16.000	105.000	.025	.228

a．Exact statistic

b．The Type IV testable hypothesis is not unique.

c．The statistic is an upper bound on F that yields a lower bound on the significance level.

d．Design: Intercept+marketid+promo+marketid * promo

Within Subjects Design: week

注释：表 5.35 中，Value 为四种检验的值，F 为统计量 *F* 值，Hypothesisdf 为样本假设检验的自由度，Error df 表示自由度的误差，Sig.为自由度的熵，Partial Eta Squared 为局部误差平方值。

表 5.35 为 SSCP 矩阵，结果包含了 4 种检验，分别 Pillai’sTrace、Wilks’Lambta、Hotelling’

Trace 和 Roy's Largest Root。

当四种检验结果不一致或不满足模型的前提假设时，Pillai'sTrace 的结果较其他统计量更稳健，检验统计量采用了 F 统计量。

检验结果提示：除 week*marketed*promo 的 Roy's Largest Root 检验外，其他均无统计学意义。但 Roy's Largest Root 的结果不可靠，下结论时应慎重。

- 重复测量单因素分析

表 5.36　重复测量单因素的分析结果

Tests of Within-Subjects Effects

Measure：sales

Source		Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
week	hericity Assumed	163.341	3	54.447	1.798	.147	.017
	Greenhouse-Geisser	163.341	2.802	58.387	1.798	.151	.017
	Huynh-Feldt	163.341	3.000	54.447	1.798	.147	.017
	Lower-bound	163.341	1.000	163.341	1.798	.183	.017
week * marketid	Sphericity Assumed	703.984[a]	27	26.073	.861	.668	.069
	Greenhouse-Geisse	703.984[a]	25.221	27.912	.861	.661	.069
	Huynh-Feldt	703.984[a]	27.000	26.073	.861	.668	.069
	Lower-bound	703.984[a]	9.000	78.220	.861	.562	.069
week * promo	Sphericity Assumed	134.762[a]	6	22.460	.742	.616	.014
	Greenhouse-Geisser	134.762[a]	5.605	24.044	.742	.607	.014
	Huynh-Feldt	134.762[a]	6.000	22.460	.742	.616	.014
	Lower-bound	134.762[a]	2.000	67.381	.742	.479	.014
week * marketid * promo	Sphericity Assumed	1624.297	48	33.840	1.118	.285	.146
	Greenhouse-Geisser	1624.297	44.838	36.226	1.118	.290	.146
	Huynh-Feldt	1624.297	48.000	33.840	1.118	.285	.146
	Lower-bound	1624.297	16.000	101.519	1.118	.349	.146
Error(week)	Sphericity Assumed	9536.287	315	30.274			
	Greenhouse-Geisser	9536.287	294.248	32.409			
	Huynh-Feldt	9536.287	315.000	30.274			
	Lower-bound	9536.287	105.000	90.822			

a．The Type IV testable hypothesis is not unique.

表 5.36 是重复测量单因素的分析结果，从表中我们可以看到校正系数在 F 统计量中发挥的作用。以 week 为例，假如模型符合前提假设，即数据满足“球对称（Sphericity Assumed）”，则自由度无需校正，仍然为 3。

再看检验结果，F 值相同，而 ρ 值不等，Greenhouse-Geisser 的结果相对保守（ρ=0.151），Lower-bound 的结果最保守（ρ=0.183）。

ρ 结果显示，无论是“week”的主效应还是其他交互效应都无统计学意义；Partial Eta Squared 的结果提示，各项对模型的贡献很小。

- 重复测量资料随时间变化的趋势

表 5.37　重复测量资料随时间变化的趋势

Tests of Within-Subjects Contrasts

Measure：sales

Source	week	Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
week	Linear	68.112	1	68.112	3.203	.076	.030
	Quadratic	32.154	1	32.154	1.021	.314	.010
	Cubic	63.075	1	63.075	1.656	.201	.016
week * marketid	Linear	123.224[a]	9	13.692	.644	.757	.052
	Quadratic	320.267[a]	9	35.585	1.130	.348	.088
	Cubic	260.493[a]	9	28.944	.760	.653	.016
week * promo	Linear	3.572[a]	2	1.786	.084	.920	.002
	Quadratic	68.715[a]	2	34.358	1.091	.339	.020
	Cubic	62.475[a]	2	31.238	.820	.443	.015
week * marketid * promo	Linear	234.148	16	14.634	.688	.800	.095
	Quadratic	400.964	16	25.060	.796	.687	.108
	Cubic	989.186	16	61.824	1.624	.075	.198
Error(week)	Linear	2232.956	105	21.266			
	Quadratic	3305.203	105	31.478			
	Cubic	3998.129	105	38.077			

a．The Type IV testable hypothesis is not unique.

表 5.37 是重复测量资料随时间的变化趋势，方法采用了多项式函数，包括线性、二阶和三阶多项式模型。

- 组间等方差性检验结果

表 5.38　组间等方差性检验结果

Levene's Test of Equality of Error Variances

	F	df1	df2	Sig.
sales.1: Units sold	1.238	27	105	.220
in thousands	1.424	27	105	.105
sales.2: Units sold	.893	27	105	.620
in thousands	1.025	27	105	.445
sales.3: Units sold				
in thousands				
sales.4: Units sold				
in thousands				

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a．

Design: Intercept+marketid+promo+marketid * promo

Within Subjects Design: week

注释：表 5.38 中：sales.1、sales.2、sales.3 和 sales.4 分别表示不同的销售方案。F 为样本统计量 *F* 值，df1 和 df2 分别表示检验统计量的自由度 1 和自由度 2，Sig.为自由度的熵。

表 5.38 提供了组间等方差性检验（方差齐性检验）结果，对于熟悉或基本熟悉统计学的读者来说，对此应不陌生，此处就不再赘述。

- 各组间效应的检验

表 5.39　各组间效应的检验

Tests of Between-Subjects Effects

Measure：sales

Transformed Variable：Average

Source	Type IV Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	888161.056	1	888161.056	35771.431	.000	.997
marketed	19721.223	9	2191.247	88.254	.000	.883
promo	5622.897	2	2811.448	113.233	.000	.683
marketid * promo	577.884	16	36.118	1.455	.131	.181
Error	2607.022	105	24.829			

a．The Type IV testable hypothesis is not unique

表 5.39 中的结果是最重要的结果。从表中可以看出，marketid 和 promo 两个因素都具有统计学意义，Partial Eta Squared 的结果显示 maketid 和 promo 对模型的贡献分别达到了 0.883 和 0.683。而两者的交互作用则无统计学意义，建议将该效应从模型中剔除。

5.6　协方差分析

协方差分析是利用线性回归的方法消除混杂因素的影响后进行的方差分析，这是实际工作中经常要考虑的问题。

例如，考虑药物对患者某个生化指标变化的影响，比较实验组与对比组的该指标变化值是否有显著差异，以确定药物的有效性；同时，可能要考虑患者病程的长短、年龄以及原指标水平对疗效的影响，要消除这些因素对药物疗效的影响，即比较实验组和对比组之间该生化指标变化量均值的差异显著性。

考虑了这些因素在测量对象的选择上使这些条件都一致时，可以使用一般的方差分析方法，这对于动物实验比较容易达到。例如，选择同品种、同一胎的大白鼠分组在相同的饲养条件下进行实验，可以相应地避免许多混杂因素的影响。但从影响因素和协变量的个数分类，协方差分析可分为单因素协方差分析、多因素协方差分析和多协变量协方差分析；从试验设计的方式分类，还可分为完全随机协方差分析、随机区组协方差分析和析因协方差分析。

【示例 5.5】政府部门对培训效果的分析。

政府为了帮助年轻人提高工作技能，进行了一系列有针对性的就业能力和工作技能培

训项目，为检验培训工作的成效，对 1000 例年轻人进行了问卷调查。表 5.40 给出了部分调查结果，主要包括培训前和培训后的收入情况。现要求对培训后的收入变量进行协方差分析。

表 5.40　1000 例工作技能培训项目调查的部分数据

年龄	性别	文化程度	婚姻状况	培训状态	培训前收入	培训后收入	待业人数
16	m	1	0	0	8.00	12.00	1
17	f	2	0	0	8.00	10.00	1
17	f	1	0	0	8.00	11.00	1
19	f	1	1	1	9.00	18.00	3
18	f	1	1	0	7.00	12.00	3
17	f	1	1	1	8.00	15.00	2
17	f	2	1	0	8.00	13.00	3
21	f	1	0	1	9.00	22.00	2
18	m	1	0	1	7.00	18.00	1
16	m	1	0	0	7.00	9.00	1
17	f	1	1	0	6.00	8.00	5
18	f	2	0	1	10.00	20.00	1
17	f	1	0	1	6.00	14.00	1
19	f	1	0	1	8.00	16.00	1
20	f	3	0	0	12.00	25.00	1
18	f	2	0	1	9.00	20.00	1
19	m	2	0	1	10.00	18.00	1
19	f	1	0	1	9.00	21.00	1
20	f	3	1	1	11.00	20.00	3
17	F	1	1	1	8.00	17.00	2
…	…	…	…	…	…	…	…

5.6.1　非饱和模型的 SPSS 操作

01 打开数据文件。

02 选择对话框。

依次单击菜单“ 分析→一般线性模型→单变量 ”命令，进入“ 单变量 ”对话框，如图 5.17 所示。在“ 因变量 ”列表框选择变量“ 培训后收入 ”；在“ 固定因子 ”列表框中选择变量“ 培训状态 ”；在“ 协方差 ”列表框中选择变量“ 培训前收入 ”。

03 设置。

单击“ 模型 ”按钮，在“ 单变量：模型 ”对话框（如图 5.18 所示）中选择“ 设定 ”选项；在“ 类型 ”下拉菜单中选择“ 主效应 ”，然后在“ 因子与协变量 ”中选择“ 培训状态 ”和“ 培训前收入 ”并引入“ 模型 ”列表框。在“ 模型 ”中显示交互效应“ 培训状态*培训

前收入 ”。单击 “ 继续 ” 按钮，返回主对话框。

单击 “ 选项 ” 按钮，在 “ 单变量：选项 ” 对话框的 “ 显示 ” 中指定输出的统计量选项，包括 “ 描述统计、功效估计、参数估计、方差齐性检验、残差图、一般估计函数 ”，如图 5.15 所示。

04 单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
UNIANOVA
    培训后收入  BY 培训状态  WITH 培训前收入
    /METHOD = SSTYPE(3)
    /INTERCEPT = INCLUDE
    /PLOT = PROFILE( 培训状态 )
    /PRINT = DESCRIPTIVE ETASQ PARAMETER HOMOGENEITY GEF
    /PLOT = RESIDUALS
    /CRITERIA = ALPHA(.05)
    /DESIGN = 培训状态 培训前收入 培训前收入*培训状态 .
```

05 单击 “ 确定 ” 按钮，执行协方差分析的操作。

协方差分析的结果解释如下：

- 组间因素

表 5.41　组间因素

Between-Subjects Factors

		N
培训状态	0	517
	1	483

表 5.41 为协方差分析的组间因素，表中列出了因素的不同组例数。

- 描述性统计量

表 5.42 显示因素在不同水平下自变量 “ 培训后收入 ” 的描述性统计量，包括均值、标准差和例数。从中可以看出，经培训后的平均收入为 18.9379，明显高于培训前收入水平的 14.4023。

表 5.42　描述性统计量

Descriptive Statistics

Dependent Variable：培训后收入

培训状态	Mean	Std. Deviation	N
0	14.4023	3.89303	517
1	18.9379	4.28162	483
Total	16.5930	4.67067	1000

- 误差方差齐性的 Levene 检验

表 5.43 给出误差方差齐性的 Levene 检验结果，F 统计量为 4.964，概率 ρ 值为 0.026<0.05，可以认为总的方差不齐。

表 5.43　误差方差齐性检验的 Levene 检验

Levene's Test of Equality of Error Variances[a]

Dependent Variable：培训后收入

F	df1	df2	Sig.
4.964	1	998	.026

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a．Design: Intercept+培训状态*培训前收入+培训状态+培训前收入

- 组间效应检验

表 5.44　组间效应检验

Tests of Between-Subjects Effects

Dependent Variable：培训后收入

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	12295.033[a]	3	4098.344	429.755	.000	.564
Intercept	131.271	1	131.271	13.765	.000	.014
培训状态*培训前收入	4.292	1	4.292	.450	.502	.000
培训状态	106.795	1	106.795	11.199	.001	.011
培训前收入	7152.586	1	7152.586	750.025	.000	.430
Error	9498.318	996	9.536			
Total	297121.000	1000				
Corrected Total	21793.351	999				

a．R Squared = .564 (Adjusted R Squared = .563)

表 5.44 给出组间效应检验结果。表中的结果表明，在组间效应的校正模型中，F 统计量为 429.755，概率 ρ 值为 0.000<0.05，可以认为培训前收入与培训后收入存在线性回归关系；而在分组变量“培训状态”中，F 统计量为 11.199，概率 ρ 值为 0.001<0.05，可以认为参加培训人员和没有参加培训人员的收入的均值不等。

交互效应项目“培训状态*培训前收入”的检验统计量 F 值为 0.450，概率 ρ 值为 0.502>0.05，表明和误差项相比，存在一个可以忽略的方差值；这意味着在因素各水平之间，可以认为协变量系数的方差齐。

- 参数估计

表 5.45 是各因素变量和交互效应的参数估计值。从表中可以看出，公共回归系数为 3.837。

表 5.45　参数估计

Parameter Estimates

Dependent Variable：培训后收入

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared
					Lower Bound	Upper Bound	
Intercept	3.837	.773	4.966	.000	2.321	5.354	.024
[培训状态=0]*	1.596	.085	18.859	.000	1.430	1.762	.263
培训前收入	1.676	.084	19.873	.000	1.510	1.841	.284
[培训状态=1]*	-3.640	1.088	-3.346	.001	-5.774	-1.505	.011
培训前收入	0[a]	.	.				
[培训状态=0]	0[a]	.	.				
[培训状态=1]							
培训前收入							

a．This parameter is set to zero because it is redundant

5.6.2　饱和模型的 SPSS 操作

前面的示例已经对培训收入的影响进行了非饱和分析，从分析中得知，模型中的交互效应项的作用不显著，这里可以进一步尝试建立饱和模型。

3．设置。

返回到“ 单变量 ”主对话框。

单击“ 模型 ”按钮，在“ 单变量：模型 ”对话框中选择“ 全因子 ”选项。单击“ 继续 ”按钮，返回主对话框。

单击“ 选项 ”按钮，在“ 单变量：按钮 ”对话框“ 显示 ”中指定输出的统计量选项，包括“ 参数估计、方差齐性检验、分布-水平图 ”。

4．单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
UNIANOVA
    培训后收入  BY 培训状态  WITH 培训前收入
    /METHOD = SSTYPE(3)
    /INTERCEPT = INCLUDE
    /PLOT = PROFILE( 培训状态 )
    /PRINT = PARAMETER HOMOGENEITY
    /PLOT = SPREADLEVEL
    /CRITERIA = ALPHA(.05)
    /DESIGN = 培训前收入 培训状态 .
```

5．单击“ 确定 ”按钮，执行协方差分析的操作。

结果解释如下：

- 误差方差齐性的 Levene 检验

表 5.46　误差方差齐性的 Levene 检验

Levene's Test of Equality of Error Variances[a]

Dependent Variable：培训后收入

F	df1	df2	Sig.
4.873	1	998	.028

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a．Design: Intercept+培训前收入+培训状态

表 5.46 是误差方差齐性检验结果，F 统计量为 4.873，概率 ρ 值为 0.028<0.05，可以认为方差相等的假设被拒绝；然后，因为由因素水平的组合自定义了两个单元，所以这个检验结果并不十分可靠。

- 组间效应检验

表 5.47　组间效应检验

Tests of Between-Subjects Effects

Dependent Variable：培训后收入

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12290.741[a]	2	6145.370	644.763	.000
Intercept	131.400	1	131.400	13.786	.000
培训前收入	7153.844	1	7153.844	750.571	.000
培训状态	4735.662	1	4735.662	496.859	.000
Error	9502.610	997	9.531		
Total	297121.00	1000			
Corrected Total	21793.351	999			

a．R Squared = .564 (Adjusted R Squared = .563)

表 5.47 是组间效应检验结果。从表中可以看出，因素变量 F 统计量的观测值为 496.859，概率 ρ 值为 0.000<0.05，所以培训前后的收入存在显著差别，可以在下面给出的参数估计表中得到具体的效应值。

- 参数估计

表 5.48　参数估计

Parameter Estimates

Dependent Variable：培训后收入

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.197	.556	7.548	.000	3.106	5.288
培训前收入	1.636	.060	27.397	.000	1.519	1.753
[培训状态=0]	-4.357	.195	-22.290	.000	-4.741	-3.974
[培训状态=1]	0[a]					

a．This parameter is set to zero because it is redundant.

表 5.48 为饱和模型的各个参数的估计值。表中显示了培训后收入的每一个预测值的效应。[培训状态=0]的参数估计值为-4.357，表示给定两个培训前具有相同收入的个体，没有参与培训的个体的收入的期望值比参加培训的个体少 4.357。

- 分布-水平图

图 5.27 为因素的两个水平之间的分布-水平图，分布-水平图显示均值和标准差的关系。同样，由于分组太少，分布-水平图没能给出更多信息。但因为两个因素水平的标准差的差距约为 0.4，小于因素水平均值之间的差距，约为 4.5，所以可以得出不同组之间具有方差齐。

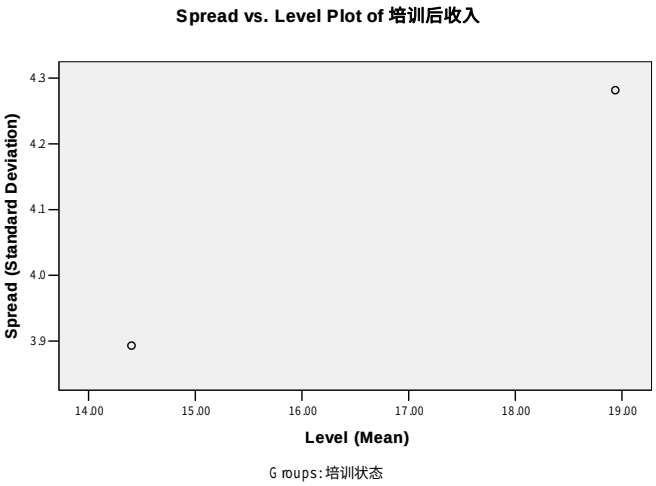


图 5.27　展布-水平图

5.7　本章小结

本章介绍了均值分析、方差分析的方法，其中方差分析由包括单因素方差分析、多因素方差分析、重复度量和协方差分析。均值分析和方差分析是专业统计分析中最重要的，同时也是最基础的统计分析方法，在多元统计中已经得到了广泛的应用，希望读者能认真学习和掌握。

第6章　参数检验

本章介绍了参数检验的方法。参数检验包括两类有密切关系的问题：参数估计和假设检验。在 SPSS 的参数检验过程中，两类问题的结果常常在输出窗口中同时给出。

SPSS 提供的“比较均值”模块可用于解决这类问题。该模块主要用于针对正态总体的均值和方差的检验和估计，包括 5 个过程：Means 过程、单样本 T 检验过程、独立样本 T 检验过程、配对样本 T 检验过程和单因素方差分析过程。单因素方差分析过程在第 5 章中已经讨论过，本章主要讨论参数检验的内容。

在 SPSS 中进行参数检验，可以利用“分析→比较均值”命令实现，菜单内容如图 6.1 所示。

本章学习要点：

- 单样本 T 检验；
- 独立样本 T 检验；
- 配对样本 T 检验。

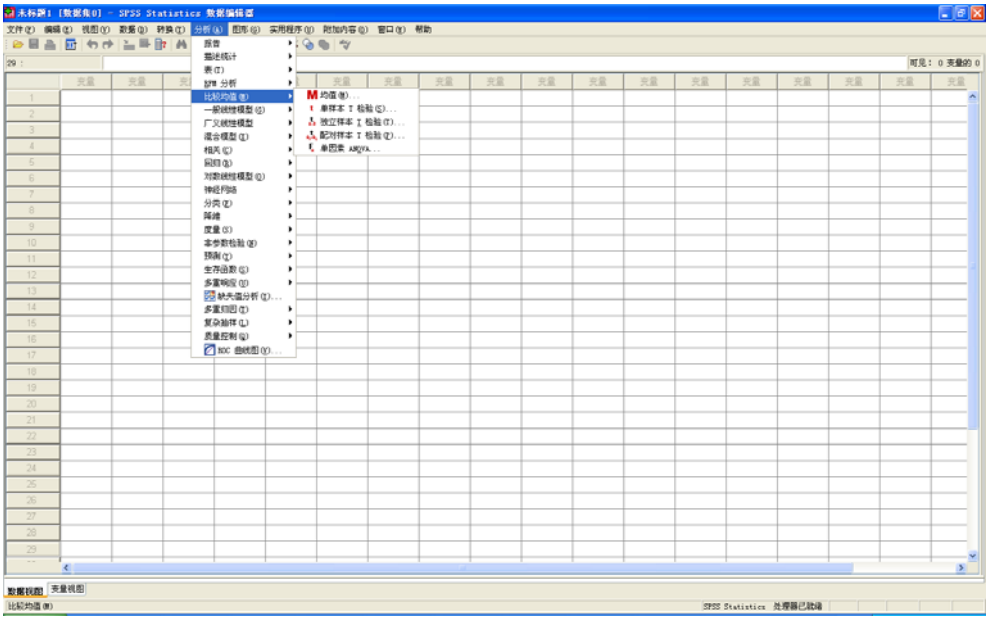


图 6.1　SPSS 比较均值模块

6.1　参数检验概述

参数检验包含两类有密切关系的问题：参数估计和假设检验。在 SPSS 的均值检验过程中，两类问题的结果常常在输出窗口中同时给出。

6.1.1　参数检验问题的提出

在实际问题的研究中，经常对研究对象的全部特征或部分特征不是很清楚，这时如果将研究对象看成一个总体，则可以从总体中抽取样本数据，利用样本数据和已知的部分信息对总体的未知特征进行统计推断。这里的未知特征可以是总体的分布，也可以是总体的某些参数，这就是统计学中的统计推断问题。

例如，某地区电视台需要对某一个节目的平均收视率进行调查，根据历史数据仅知道某一确定时刻收看该节目的人数服从正态分布，但不知道平均收看该节目的人数为多少。对于这类问题的研究，电视台很难对该地区的每一个观众进行逐一调查，只能采用抽样技术，从所有居民中随机选取一部分观众进行调查，从而推断总体均值的大小。

再如，图书馆需要对每天数据的借阅量进行推断，但由于没有历史数据，无法知道作为总体的借阅量的概率分布，更无法知道借阅量的平均值和平均偏差，在这种情况下，总体数据是无法收集到的，因为借阅量的全部数据还包括没有发生的未来借阅量，此时只能通过抽取有限的样本数据，去推断总体的分布以及相关参数。

以上两个例子代表统计推断的两种主要方法。

一，在总体分布已知的情况下，对总体的参数进行推断。例如，当已知总体服从正态分布时，需要对正态分布的均值和方差进行推断，在一定的置信度条件下估计参数的取值范围，或者在一定的显著性水平下，对给定参数的取值进行检验。一般来说，上述总体已知情况下的统计推断问题称之为参数检验问题。参数检验不仅可以针对一个总体的检验，也可以针对两个或多个总体的比较问题。例如，上例中关于电视台的平均收视率的推断问题可以归结为参数检验问题。

二，当总体分布未知时，需要根据样本推断总体的分布类型和参数值的大小，这类总体分布未知的统计推断过程称为非参数检验。例如，上例中关于图书馆借阅量的分布问题就是一个非参数检验问题。此类问题将专门在非参数检验一节详细阐述。

6.1.2 参数检验的基本步骤

对正态总体参数的检验过程一般包括参数的假设检验和参数估计。假设检验是对给定的总体参数值，利用样本数据对其推断，并给出接受或者拒绝的过程。假设检验的基本原则是依据统计推断原理，即小概率事件在一次特定的抽样中一般是不会发生的，如果发生了小概率事件，我们有理由怀疑假设的正确性，从而拒绝检验该问题时做出的假设。在具体操作中，首先应定义所谓的小概率，一般取 0.01 或 0.05，即显著性水平。显著性水平取得太小，容易发生取伪错误；取得太大，则容易发生弃真错误。

例如，某一个学校学生的高等数学成绩服从正态分布，根据往年数据得知其均值为 75 分。为检验今年新生的高等数学平均成绩和往年相比是否有显著差异，可以随机抽取 20 名新生的高等数学成绩。显然，这 20 名新生的数学成绩平均成绩与 75 分越接近，这批新生的高等数学成绩与往届学生无差异的可能越大。由于是统计抽样得到的数据，当样本的平均成绩和 75 分很接近时，我们只能保证总体的均值为 75 分的概率很大，所以，参数检验的结论是在概率意义下成立的。

根据上述讨论，可归纳出参数检验的基本步骤。

01 给出检验问题的零假设。

根据检验问题的要求，将需要检验的最终结果作为零假设。例如，在上例中，需要检验平均成绩是否为 75 分，由此可做出零假设，$H_0:\mu=75$。

02 选择检验统计量。

在统计推断中，总是通过构造样本的统计量并计算该统计量的概率值进行推断，一般构造的统计量应该服从或近似服从常用的已知分布，例如均值检验中最常用的 t 分布和 F 分布等。

03 计算检验统计量的观测值及其发生的概率值。

在给定零假设的前提下，计算统计量的观测值和响应概率 ρ 值。概率 ρ 值就是在零假设 H_0 成立时检验统计量的观测值发生的概率，该概率值间接地给出了样本值在零假设成

立前提下的概率，对此可以依据意义的标准来判定其发生的概率是否为小概率。

04 在给定显著性水平的条件下，做出统计推断结果。

这里的显著性水平是指当假设正确时被拒绝的概率，即弃真概率，一般取 0.01 或 0.05。当检验统计量的概率 ρ 值小于显著性水平时，则认为此时拒绝零假设而犯弃真错误的概率小于显著性水平，即低于预先给定的水平，也就是说犯错误的概率小到我们能容忍的范围，这时可以拒绝零假设；反之，如果检验统计量的概率 ρ 值大于显著性水平，如果拒绝零假设，犯弃真错误的概率大于预先给定的容忍水平，这时不应该拒绝零假设。

所以，在 SPSS 的检验问题中，都是利用概率 ρ 值和显著性水平来进行比较而做出拒绝或者接受零假设结论的；SPSS 中系统将自动计算概率 ρ 值，但显著性水平应该由用户事先设定。

在检验过程中，不仅可以计算出统计量的概率 ρ 值，SPSS 还可以给出在一定的置信度下，检验量的置信区间，这是属于统计分析中参数估计的内容。参数估计中构造的统计量和统一参数检验中的统计量是相同的，这便是可以同时给出假设检验和参数估计结果的关键所在。

6.2　单样本 T 检验

单样本 T 检验用于检验正态总体的均值是否与给定的检验值之间存在显著性差异。要求样本数据来自于服从正态分布的单总体，这种检验之所以称为 T 检验，是因为检验过程中构造的统计量服从 t 分布，也称为单样本的均值检验。

6.2.1　单样本 T 检验的基本方法

在实际问题研究中，单样本均值检验的例子很多。例如，上例中推断新生的平均成绩是否为 75 分；在人口普查中，某地区职工今年的平均收入是否和往年的平均收入有显著差异。

在给定样本来自于正态总体的假设下，可以利用假设检验基本方法进行均值的检验，具体步骤和一般的参数检验步骤完全一致，在下面给出的步骤中，重点讨论单样本均值检验统计量的构造。

01 给出均值检验的零假设。

单样本 T 检验需要检验总体均值与指定的检验值是否存在显著性差异。为此，给出检验均值 μ_0，零假设 $\mu=\mu_0$，这里 μ 为总体均值。

例如，检验平均成绩是否为 75 分，其零假设 $H_0:\mu=75$。

02 选择检验统计量。

均值检验的前提是总体服从正态分布 $N(\mu,\delta^2)$，这里 μ 为总体均值，δ^2 为总体方差。如果样本容量为 n，样本均值为 $\bar{X}$，则 $\bar{X}$ 仍服从正态分布，即：

$$\bar{X} \sim N(\mu, \frac{\delta^2}{n}) \tag{6.1}$$

在零假设成立的条件下，均值检验使用 t 统计量，构造的 t 统计量为：

$$t = \frac{\bar{X} - \mu}{S/\sqrt{\mu}} \tag{6.2}$$

这里，当 $\mu = \mu_0$ 时，t 统计量服从自由度为 $n-1$ 的 t 分布，S 为样本标准差，定义为：

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{6.3}$$

SPSS 的操作结果中还显示均值标准误差（Std.Error Mean），计算公式为：$\frac{S}{\sqrt{n}}$，即 t 统计量的分母部分。

- 计算检验统计量的观测值及其发生的概率

在给定零假设的前提下，SPSS将检验值μ_0代入t统计量，得到检验统计量的观测值以及根据 t 分布的分布函数计算出的概率 ρ 值。

- 给定显著性水平，做出统计推断结果

当检验统计量的概率 ρ 值小于显著性水平时，则拒绝零假设，认为总体均值与μ_0之间存在显著差异；反之，如果检验统计量的概率 ρ 值大于显著性水平，则接受零假设，认为总体均值与μ_0之间无显著性差异。

6.2.2 单样本 T 检验的实例

SPSS 中单样本 T 检验的操作步骤如下。

01 打开检验对话框。

依次单击菜单“分析→比较均值→单样本 T 检验”命令，进入“单样本 T 检验”对话框，如图 6.2 所示。



图 6.2 “单样本 T 检验”对话框

02 选择检验变量。

在“单样本 T 检验”对话框左侧的变量类表框中选择进行检验的变量，单击 按钮，

将其移动到右侧的“检验变量”列表框中；并在“检验值”文本框中输入检验值。

03 设置“选项”对话框及选项。

单击“选项”按钮，弹出“单样本 T 检验：选项”对话框，如图 6.3 所示，该对话框用于指定置信区间和缺失值的处理方法。

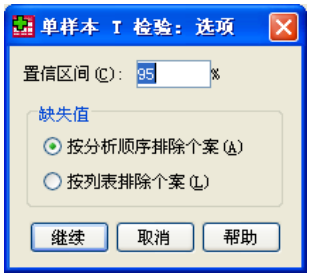


图 6.3 “单样本 T 检验：选项”对话框

在“置信区间”文本框中输入置信水平，默认置信水平为 95%，也可以输入 1~99 之间的任意数字。在“缺失值”选项组中指定缺失值的处理方式，有两个单选项，分别为：“按分析顺序排除个案”和“按列表排除个案”。

设置完毕后，或单击“继续”按钮，确认操作，返回主对话框；也可以单击“取消”按钮，删除操作；单击“帮助”按钮，进行在线帮助操作。

04 执行操作。

在“单样本 T 检验”主对话框中，单击“确定”按钮，完成操作。

【示例 6.1】圆盘制动闸直径的T检验。

一个生产高性能汽车的公司生产直径为 322mm 的圆盘制动闸。公司的质量控制部门随机抽取不同机器生产的制动闸进行检验。现一共有 4 台机器，每台机器抽取 16 个产品。表 6.1 给出测量结果。

表 6.1　不同机器生产的圆盘制动闸直径测量数据

机器编号	圆盘制动闸直径
机器 1	322.0003 322.0048 322.0215 321.9907 322.0109 321.9954 322.0059 321.9759 321.9981 321.9957 321.9841 321.9836 322.0037 322.0000 322.0030 322.0023
机器 2	322.0069 322.0306 322.0114 322.0289 322.0087 322.0257 322.0176 322.0070 322.0183 321.9864 322.0178 322.0182 322.0200 322.0122 322.0137 322.0048
机器 3	321.9863 321.9899 322.0016 321.9843 322.0174 321.9834 322.0024 322.0005 322.0039 322.0161 322.0017 322.0085 321.9902 321.9932 321.9912 322.0019
机器 4	321.9939 322.0030 322.0055 322.0028 321.9863 322.0017 321.9978 321.9914 321.9957 321.9989 321.9898 322.0023 321.9861 321.9910 321.9825 321.9982

利用单样本 T 检验方法检验每个机器生产的产品均值和 322 在 90%的置信水平下是否有显著差异。

由于该问题设计的是 4 个单总体，而且要进行总体均值检验，同一机器生产的产品直径可以近似认为服从正态分布；因此，可以利用 SPSS 的单样本 T 检验过程进行分析。

SPSS 单样本检验的操作步骤如下。

01 建立数据文件。

数据文件包含两个变量，分别为“制动闸直径”和“机器编号”，其中变量“机器编号”为定类型变量，用于记录机器的序号，变量值为 1~4，如图 6.2 所示。

02 拆分数据文件。

机器序号为定类型变量，区别不同机器生产的制动闸。由于是对不同机器生产的产品分别进行检验，所以必须先利用变量“机器编号”对数据文件进行拆分。

依次单击菜单“数据→分割文件”命令，在弹出的“分割文件”对话框（如图 6.4 所示）中，选择“比较组”，再将左侧变量列表框中的变量“机器编号”移动至右侧“分组方式”列表框中，单击“确定”按钮，执行拆分操作。



图 6.4 “拆分文件”对话框

03 选择检验对话框。

依次单击菜单“分析→比较均值→单样本 T 检验”命令，进入“单样本 T 检验”对话框。在“检验值”文本框上输入 322。如图 6.2 所示。

04 设置“选项”对话框。

单击“选项”按钮，进入“单样本 T 检验：选项”对话框，在“置信区间”文本框中输入数值 90；单击“继续”按钮，返回主对话框。

05 执行操作。

在“单样本 T 检验”对话框单击“确定”按钮，执行单样本 T 检验的操作。

06 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
T-TEST
    /TESTVAL = 322
```

```
/MISSING = ANALYSIS
/VARIABLES = 制动闸直径
/CRITERIA = CI(.9) .
```

执行完“单样本 T 检验”后，得到单样本 T 检验的结果。

- 单样本 T 检验的描述性统计量

表 6.2　单样本 T 检验的描述性统计量

One-Sample Statistics

机器编号		N	Mean	Std.Deviation	Std.Error Mean
1	制动闸直径	16	321.9985	.0111445	.0027861
2	制动闸直径	16	322.0143	.0106985	.0026746
3	制动闸直径	16	321.9983	.0104840	.0026210
4	制动闸直径	16	321.9954	.0069878	.0017469

表 6.2 给出了单样本 T 检验的描述性统计量，包括每一个样本的均值、标准差和均值标准误差，均值标准误差的计算是标准差和样本容量开方的比值。

- 单样本 T 检验的检验结果

表 6.3 是单样本 T 检验的结果列表。表中第二列是 t 统计量的观测值，分别为-0.541、5.333、-0.656 和-2.615；第三列 df 表示自由度，t 统计量的自由度为样本容量减 1，每一个样本的容量都为 15，故自由度为 15（16-1）；第四列 $Sig.(2-tailed)$ 表示统计量的双尾概率 ρ 值；第五列是样本均值和检验值的差；第六列和第七列分别是在置信度为 90%时总体均值与原检验值之差的置信下限和置信上限，共同构成了该差值的置信区间，例如第一组样本的置信区间为（-0.006390 0.003378）。

表 6.3　单样本 T 检验的结果列表

One-Sample Test

机器编号		Test Value=322					
		t	df	Sig.（2-tailed）	Mean Difference	90% Confidence Interval of the Difference	
						Lower	Upper
1	制动闸直径	-.541	15	.597	-.0015063	-.006390	.003378
2	制动闸直径	5.333	15	.000	.0142625	.009574	.018951
3	制动闸直径	-.656	15	.522	-.0017187	-.006313	.002876
4	制动闸直径	-2.615	15	.019	-.0045688	-.007631	-.001506

置信度为 90%时，显著性水平为 0.10，第一组的双尾概率 ρ 值为 0.597，大于 0.10，故零假设成立；观察其他样本的概率 ρ 值可以发现，除第三组的概率 ρ 值小于 0.10 以外，其余都大于 0.10，故可以认为编号为 3 的机器生产的产品均值与 322 有显著差异，其余都可以认为产品均值为 322。

6.3 独立样本 T 检验

两独立样本的 T 检验用于检验两个独立样本是否来自于具有相同均值的总体，也就是检验两个独立正态总体的均值是否相等。

例如，比较两所高校学生的英语成绩。假设两所学校的英语成绩是相互独立的，需要分析两所学校的平均成绩是否有显著差异。

在上例中，两所学校的成绩可以看成两个独立总体，一般可假设英语成绩服从正态分布，检验统计量仍为 t 分布，所以这种检验类型也称为两独立样本的均值 T 检验。

6.3.1 独立样本 T 检验的基本方法

在满足两个样本均来自正态总体的假设下，可以利用假设检验基本方法进行均值的检验，具体步骤和一般的参数检验步骤完全一致，在下面给出的步骤中将重点讨论两独立样本均值检验统计量的构造。

01 提出零假设。

独立样本 T 检验需要检验两个总体的均值是否存在显著性差异。其零假设为 $H_0:\mu_1-\mu_2=0$，这里 μ_1、μ_2 分别为两总体的均值。

例如，检验两个学校的英语成绩是否相等，其零假设为 $H_0:\mu_1-\mu_2=0$。

02 选择检验统计量。

两独立样本均值检验的前提是，两个独立的总体分别服从 $N(\mu_1,\delta_1^2)$ 和 $N(\mu_2,\delta_2^2)$，这里 δ_1^2、δ_2^2 分别为两总体的方差。

在零假设成立的条件下，两独立样本的均值检验使用 t 统计量。构造两独立样本的 t 统计量必须考虑两种情况。

1．当两个总体方差未知但相等，即 $\delta_1^2=\delta_2^2$ 时，构造的 t 检验统计量定义为：

$$t=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{6.4}$$

这里，n_1 和 n_2 分别为两样本容量；S_1 和 S_2 分别为两样本标准差，且 $S_\omega=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$，并且该统计量服从自由度为 n_1+n_2-2 的t分布。

2．当两个总体方差未知且不相等，即 $\delta_1^2\neq\delta_2^2$ 时，构造的 t 检验统计量定义为：

$$t=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \tag{6.5}$$

可以证明，该检验统计量服从修正自由度的 t 分布，其修正自由度为：

$$df=\frac{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1}+\frac{(\frac{S_2^2}{n_2})^2}{n_2}} \tag{6.6}$$

在统计分析中，如果两个总体的方差相等，则称之为满足方差齐性。确定两个独立样本的方差齐性是构造和选择独立样本 T 检验统计量的关键。SPSS 利用 Levene F 方差齐性检验方法检验两总体方差是否存在显著变异。

进行 Levene F 方差齐性检验，首先提出零假设 $H_0:\delta_1^2=\delta_2^2$；执行检验过程中，若概率 ρ 值小于给定的显著性水平（一般为 0.05），则拒绝零假设 H_0，认为两个总体的方差不等；否则认为两个总体的方差无显著差异。

F 检验中计算 F 统计量值的公式为：

$$F=\frac{\max(S_1^2,S_2^2)}{\min(S_1^2,S_2^2)}\sim F(n_1-1,n_2-1) \tag{6.7}$$

这里 n_1-1 为 $\max(S_1^2,S_2^2)$ 的自由度，n_2-1 为 $\min(S_1^2,S_2^2)$ 的自由度。

3．计算检验统计量的观测值及其发生的概率。

在给定零假设前提下，SPSS 将检验值 0 代入 t 统计量的 $\mu_1-\mu_2$ 部分，得到检验统计量的观测值以及根据 t 分布的分布函数计算出概率 ρ 值。

4．给定显著性水平，做出统计推断结果。

当检验统计量的概率 ρ 值小于显著性水平时，则拒绝零假设，认为总体均值与 μ_0 存在显著差异；反之，如果检验统计量的概率 ρ 值大于显著性水平时，则接受零假设，认为这两个总体均值无显著差异。

6.3.2　独立样本 T 检验的实例

SPSS 中独立样本 T 检验的操作步骤如下。

01 打开 T 检验对话框。

依次单击菜单“分析→比较均值→独立样本 T 检验”，进入“独立样本 T 检验”对话框，如图 6.5 所示。

在左侧的变量列表框中选择进行检验的变量，单击 按钮，将其移动到右侧的“检验变量”列表框中。

02 选择分组变量和分界点。

在左侧的变量列表框中选择作为分组标识的变量，移动到“分组变量”文本框中，然后单击“定义组”按钮，显示如图 6.6 所示的“定义组”对话框。

在该对话框中，分别在“组 1”和“组 2”文本框中输入作为第一组和第二组样本的分组变量值；如果分组变量是连续型的定距型变量，则可以在“割点”文本框中输入分界点，分界点把数据分为两部分，大于该值的数据为一组，小于此值的为另一组。单击“继

续”按钮，返回“独立样本 T 检验”对话框。



图 6.5 “独立样本 T 检验”对话框

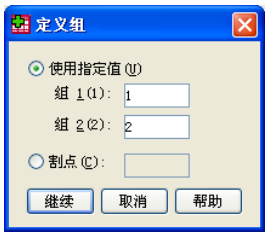


图 6.6 “定义组”对话框

03 设置“选项”对话框及其选项。

单击“选项”按钮，出现“独立样本 T 检验：选项”对话框，如图 6.7 所示，在对话框中设置用于指定置信水平和缺失值的处理方法。

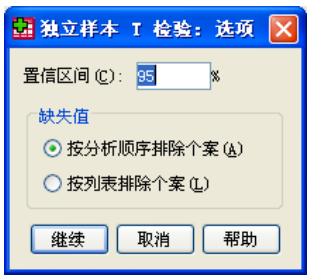


图 6.7 “独立样本 T 检验：选项”对话框

在“置信区间”文本框中输入置信水平，默认置信水平为 95%，也可以输入 1~99 之间的任意数字。“缺失值”选项组指定缺失值的处理方式，有两个选项，分别是“按分析顺序排除个案”和“按列表排除个案”。

设置完毕后，单击“继续”按钮，确认操作，返回主对话框；也可以单击“取消”按钮，删除操作；单击“帮助”按钮，进行在线帮助操作。

04 执行操作。

在“独立样本 T 检验”主对话框中，单击“确定”按钮，执行操作。

【示例 6.2】两组乒乓球得分数据的T检验。

在体育课上记录 14 名学生乒乓球得分的数据，男女各 7 名，他们的得分数据如表 6.4 所示。

比较在置信度为 95%的情况下男女得分时的明显差别。

1．建立数据文件。

数据文件中包含两个变量，变量“得分”记录学生的成绩；变量“性别”记录学生的性别，其中 1 表示男生，2 表示女生。

表 6.4　体育课上 14 名学生乒乓球得分数据

性　别	得　分
男	82.00 80.00 85.00 85.00 78.00 87.00 82.00
女	75.00 76.00 80.00 77.00 80.00 77.00 73.00

2．选择检验对话框。

依次选择菜单“分析→比较均值→独立样本 T 检验”命令，进入“独立样本 T 检验”主对话框。在该对话框中，选择变量“得分”，将其移动到“检验变量”列表框；将变量“性别”移动到“分组变量”列表框中，作为分组变量，然后单击“定义组”按钮，弹出“定义组”对话框，在“组 1”和“组 2”文本框中分别输入“1”和“2”；单击“继续”按钮，返回到主对话框。

3．单击“确定”按钮，执行操作。

4．单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
T-TEST
    GROUPS = 性别(1 2)
    /MISSING = ANALYSIS
    /VARIABLES = 得分
    /CRITERIA = CI(.95) .
```

执行完“独立样本 T 检验”后，得到独立样本 T 检验的结果。

- 独立样本 T 检验的基本描述性统计量

表 6.5　独立样本 T 检验的基本描述性统计量

Group Statistics

性　别		N	Mean	Std Deviation	Std.Error Mean
得分	男	7	82.7143	3.14718	1.18952
	女	7	76.8571	2.54484	.96186

表 6.5 给出的是该问题的独立样本 T 检验的描述性统计量，包括两个样本的均值、标准差和均值标准误差。从表中可以看出，男生的平均成绩为 82.7143，女生的平均成绩为 76.8571，两者之间有一定的差距。

- 独立样本 T 检验的检验结果

表 6.6 是独立样本 T 检验的均值检验结果。第二列是检验 F 统计量的观察值，为 0.622；第三列为对应的概率 ρ 值，为 0.445。如果显著性水平为 0.05，由于概率 ρ 值大于 0.05，可以认为两个总体的方差无明显差异，即通过了 Levene 方差齐性检验。

由于两个总体的方差无显著差异，所以 T 检验结果应在方差相等的情况下做出，故推断结果应从“Equal variances assumed”行中得到，在该行的第四列给出 t 统计量的观察值，为 3.829，第五列为相应的双尾概率 ρ 值，为 0.002；在显著性水平为 0.05 的情况下，t 统计量的概率 ρ 值小于 0.05，故拒绝零假设，因此认为两个总体的均值存在显著差异，即男女生的体育成绩平均得分存在显著差异。

表 6.6　独立样本 T 检验的检验结果

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig	t	df	Sig（2-tailed）	Mean Difference	Std.Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
得分　Epual variancex assumed	.622	.445	3.829	12	.002	5.85714	1.52975	2.52410	9.19018
Equal variances not assumed			3.829	11.496	.003	5.85714	1.52975	2.50784	9.20645

表中第五列为统计量的自由度，为 12（7+7-2）；第七列和第八列分别为 *t* 统计量的两个总体均值差的均值和两个总体均值差的标准误差；第九列和第十列为置信度为 95%时的两个总体均值的置信下限和置信上限，可以得到两个总体均值差的置信度为 95%的置信区间为（2.50784 2.52410），该置信区间不包括零值，进一步证实了上述推断。

6.4　配对样本 T 检验

配对样本的 T 检验用于检验两个相关样本是否来自具有相同均值的总体，即对于两个配对样本，推断两个总体的均值是否存在显著差异。

配对的概念是指两个样本的各样本值之间存在着对应关系。例如，分析一组病人治疗前和治疗后的体重变化，显然，是同一个病人对应治疗前后两种不同的体重，同一个病人的两种体重就是一个配对，若干个病人治疗前后两组体重构成了配对样本。从前面的描述可以看出，配对样本的两个样本值之间的配对是一一对应的，并且两个样本的容量相同。

配对样本的 T 检验是求出每个个案值之差，所有样本值的观测值之差形成一个新的单样本。显然，如果两个样本的均值没有显著差异，则样本值之差的均值应该接近零，这实际上转换成为一个单样本的 T 检验。所以，配对样本的 T 检验就是检验来自总体均值的差值是否为零，这里要求差值来自总体服从正态分布。

6.4.1　配对样本 T 检验的基本方法

在满足两个样本差值来自正态总体分布的假设下，可以利用假设检验基本方法进行均值的检验，具体步骤和一般的参数检验步骤完全一致，下面给出的步骤将重点讨论配对样本 T 检验的统计量的构造。

01 提出零假设。

配对样本 T 检验需要检验两个总体的均值是否存在显著性差异。其零假设为 $H_0:\mu_1-\mu_2=0$，这里 μ_1、μ_2 分别为两总体的均值。

02 选择检验统计量。

设 $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ 为配对样本，差值 $d_i=X_i-Y_i,i=1,2,...,n$。配对样本 T 检验的前提是两个样本的差值来自总体 d 服从正态分布。

在零假设成立的条件下，差值来自总体 d 的均值为零，配对 T 检验使用 t 统计量，构造 t 统计量为：

$$t=\frac{\bar{d}-(\mu_1-\mu_2)}{S/\sqrt{n}} \tag{6.8}$$

这里，当 $\mu_1-\mu_2=0$ 时，t 统计量服从自由度为 $n-1$ 的 t 分布。

03 计算检验统计量的观测值及其发生的概率。

在给定零假设的前提下，SPSS 将检验值 μ_0 代入 t 统计量，得到检验统计量的观测值以及根据 t 分布函数计算出的概率 ρ 值。

04 给定显著性水平，做出统计推断结果。

当检验统计量的概率 ρ 值小于显著性水平时，则拒绝零假设，认为两个总体的均值存在显著性差异；反之，如果检验统计量的概率 ρ 值大于显著性水平，则接受零假设，认为两个总体的均值无显著差异。

6.4.2　配对样本 T 检验的实例

SPSS 中两配对样本 T 检验的操作步骤如下。

01 选择对话框和检验变量。

依次单击菜单“分析→比较均值→配对样本 T 检验”命令，进入“配对样本 T 检验”对话框，如图 6.8 所示。

在“配对样本 T 检验”对话框左侧的变量列表框中选择成对变量，单击 按钮，将其选入到右侧的“成对变量”列表框，可以选择多对检验样本。

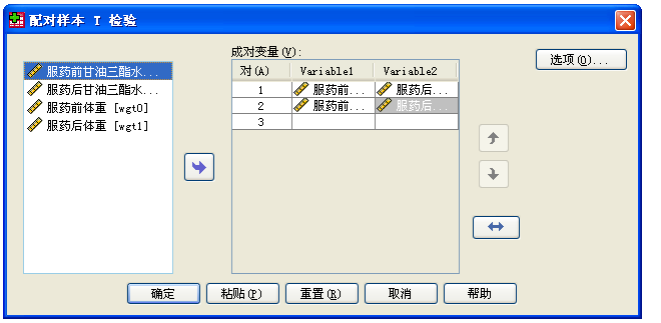


图 6.8　“配对样本 T 检验”对话框

02 设置“选项”对话框及其选项。

单击“选项”按钮，出现“配对样本 T 检验：选项”对话框，如图 6.9 所示。该对话框用于指定置信水平和缺失值的处理方法，与单样本 T 检验的“选项”对话框完全相同。

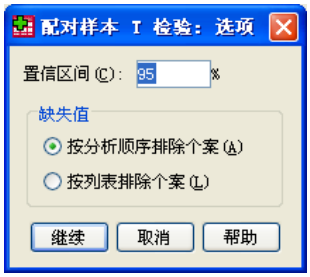


图 6.9 “配对样本 T 检验：选项”对话框

在“置信区间”文本框中输入置信水平，默认置信水平为 95%，也可以输入 1~99 之间的任意数字。“缺失值”选项组指定缺失值的处理方式，有两个选项分别是“按分析顺序排除个案”和“按列表排除个案”。

设置完毕后，单击“继续”按钮，确认操作，返回主对话框；也可以单击“取消”按钮，删除操作；单击“帮助”按钮，进行在线帮助操作。

03 执行操作。

在“独立样本 T 检验”主对话框中，单击“确定”按钮，执行操作。

【示例 6.3】新药疗效的测试结果检验。

某医疗机构针对具有家族心脏病史的病人研发了一种新药。为了检验这种新药的疗效是否显著，对 16 位病人进行了为期半年的观察测试，测试指标为使用该药之前和之后的体重以及甘油三酯的水平的变化。表 6.7 给出了本例数据。

表 6.7 服药前后甘油三酯水平和体重数据

	服药前后甘油三酯水平和体重数据
服药前甘油三酯水平	180.00 139.00 152.00 112.00 156.00 167.00 138.00 160.00 107.00 156.00 94.00 10.700 145.00 186.00 112.00 104.00
服药后甘油三酯水平	100.00 92.00 118.00 82.00 97.00 171.00 132.00 123.00 174.00 92.00 121.00 150.00 159.00 101.00 148.00 130.00
服药前体重	198.00 237.00 233.00 179.00 219.00 169.00 222.00 167.00 199.00 233.00 179.00 158.00 157.00 216.00 257.00 151.00
服药后体重	192.00 225.00 226.00 172.00 214.00 161.00 210.00 161.00 193.00 226.00 173.00 154.00 143.00 206.00 249.00 140.00

可以认为：甘油三酯水平和体重都近似服从正态分布，数据是成对出现的，所以可以利用 SPSS 的两配对样本的 T 检验过程。

该问题的具体操作步骤如下。

01 建立数据文件。

数据操作文件包含 4 个变量（如图 6.8 所示），“tg0”和“tg1”表示服药前后的甘油三酯水平，“wgt0”和“wgt1”表示服药前后的体重。

02 选择检验对话框。

依次单击菜单“分析→比较均值→配对样本 T 检验”，进入“配对样本 T 检验”对话框，选择“tg0”和“tg1”为第一对检验变量，将其移到“配对变量”列表框中；类似地，再将变量“wgt0”和“wgt1”移至“配对变量”列表框。如图 6.8 所示。

03 单击“确定”按钮，执行操作。

04 单击“粘贴”按钮得到的 Syntax 命令语句程序如下：

```
T-TEST
    PAIRS = tg0--tg1 wgt0--wgt1  WITH  _ (PAIRED)
    /CRITERIA = CI(.95)
    /MISSING = ANALYSIS.
```

执行完“配对样本 T 检验”后，得到配对样本 T 检验的结果。

- 两配对样本 T 检验的基本描述统计量

表 6.8　两配对样本 T 检验的基本描述性统计量

Paired Samples Statistics

		Mean	N	Std.Deviation	Std.Error Mean
Pair 1	服药前甘油三酯水平	138.44	16	29.040	7.260
	服药后甘油三酯水平	124.38	16	29.412	7.353
Pair 2	服药前体重	198.38	16	33.472	8.368
	服药后体重	190.31	16	33.508	8.377

表 6.8 是两配对样本 T 检验的基本描述性统计量，包括每一个样本的均值、标准差和均值标准误差。从两对样本的均值变化可以看出，均值都应有一定量的变化，但是否存在显著性差异，还必须通过计算相应的 t 统计量来判断。

- 两配对样本 T 检验的简单相关关系及其检验

表 6.9　两配对样本相关性检验结果

Paired Samples Correlations

		N	Correlation	Sig
Pair 1	服药前甘油三酯水平&服药后甘油三酯水平	16	-.286	.283
Pair 2	服药前体重&服药后体重	16	.996	.000

表 6.9 是两配对样本 T 检验的简单相关关系及其检验结果。表中第三列为服药前后两样本的相关系数，第四列是相关系数的检验 ρ 值。

从表中可以看出，在显著性水平为 0.05 时，甘油三酯前后水平的概率 ρ 值为 0.283，

大于 0.05，接受零假设，可以认为服药前后的甘油三酯水平没有明显的线性关系；但注意到，服药前后体重的相关系数概率 ρ 值为 0.000，小于 0.05，决绝了零检验，可以认为服药前后的体重具有一定的线性关系。

- 两配对样本 T 检验的结果

表 6.10　两配对样本 T 检验结果

Paried Samples Text

		Paired Differences					t	df	Sig（2-tailed）
		Mean	Std Deviation	Std.Error Mean	95% Confidence Inferval of the Difference				
					Lower	Upper			
Pair 1	服药前甘油三酯水平-服药后甘油三酯水平	14.063	46.875	11.719	-10.915	39.040	1.200	15	.249
Pair 2	服药前体重-服药后体重	8.063	2.886	.722	6.525	9.600	11.175	15	.000

表 6.10 是两配对样本 T 检验的最终结果。表中第二列是配对样本的平均差值，服药前后的体重平均相差了 8.063；而甘油三酯水平相差了 14.063；第三列是差值的标准差；第四列是差值的均值标准误差；第五列和第六列分别是在置信度为 95%时差值的置信下限和置信上限，共同构成了该差值的置信区间，例如第一对样本的总体差值置信区间为（-10.915　39.040），而第二对样本的总体差值的置信区间为（6.525　9.600），0 不包括在第二对的置信区间，第二对样本的差异较大。

第七列为统计量的观测值；第八列为自由度；最后一列为双尾检验概率 ρ 值，在置信水平为 95%时，显著性水平为 0.05，第一组的双尾概率 ρ 值为 0.249，大于 0.05，故零假设成立，可以认为服药前后甘油三酯水平没有显著差异，即服药对降低甘油三酯水平没有效果。

第二组的双尾概率 ρ 值为 0.000，小于 0.05，拒绝零假设，可以认为服药前后的平均体重有显著差异，即服药对降低体重有明显效果。

6.5　本章小结

本章讨论了统计分析中的参数检验方法。参数检验问题一般分为三个部分：单样本 T 检验、独立样本 T 检验和配对样本 T 检验，处理方式各有各的特点，也不尽相同。读者学习时，应重点掌握参数检验的基本思想以及统计结果输出的实际含义。

第7章 非参数检验

非参数检验是指总体分布未知或知之甚少时，利用已知样本数据对总体的分布形态等做出推断的分析方法，非参数检验方法在推断过程中一般不涉及总体参数。非参数检验是相对于参数检验而言的，这两种检验方法在实际生活中有着广泛的应用，并且共同构成了统计推断的基本内容。例如，超市柜台某一时刻的排队人数是否服从泊松分布，可以通过一定量的观测数据，利用非参数检验的检验方法做出假设，给出推断过程和结果。

和参数检验方法一样，非参数检验过程也是先根据问题做出假设，然后利用统计学原理构造出适当的统计量，最后利用样本数据计算统计量的概率 ρ 值，与显著性水平比较，得出拒绝或者接受假设的结果。

非参数检验的方法非常丰富。根据所要推断的问题或者所利用的统计量，可以分为卡方检验、二项分布检验、游程检验、单样本 K-S 检验、两个独立样本检验、K 个独立样本检验、两个相关样本检验和 K 个相关样本检验等。在 SPSS 中，可以利用“分析→非参数检验”命令实现非参数检验，菜单内容如图 7.1 所示。

本章学习要点：

- 卡方检验；
- 二项分布检验；
- 游程检验；
- 单样本 K-S 检验；
- 两个独立样本检验；
- K 个独立样本检验；
- 两个相关样本检验；
- K 个相关样本检验。

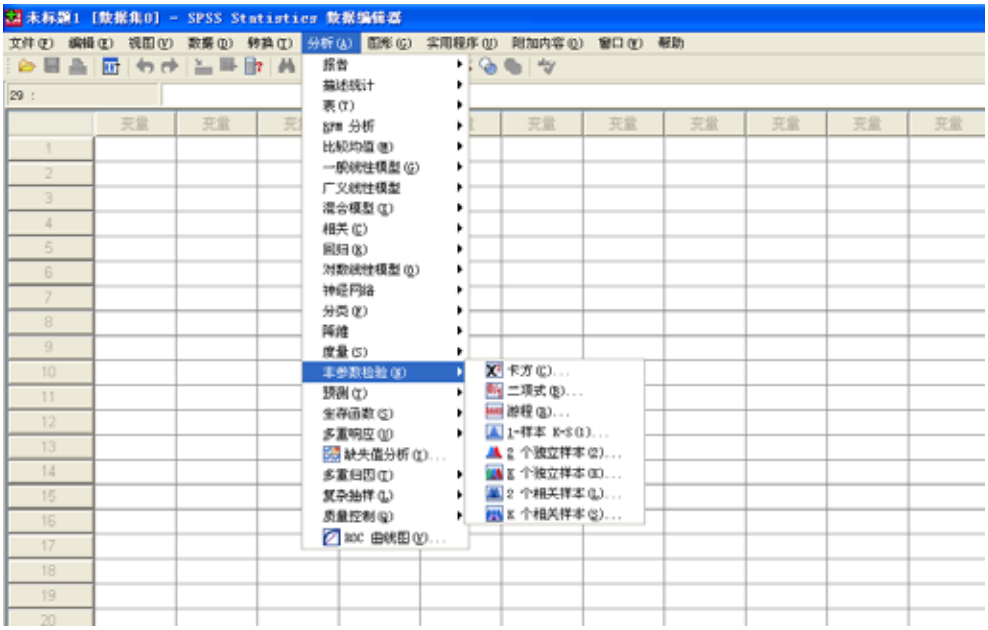


图 7.1　SPSS 非参数检验模块

7.1　卡方检验

卡方检验（Chi-Square Test）法也称卡方拟合优度检验，用于检验样本数据是否与某种概率分布的理论数值相符合，进而推断样本数据是否来自于该分布的样本问题。例如，根据摇奖号码中出现的数字检验摇奖的球是否均匀，即 0~9 每个数字出现的概率是否均是 1/10。

7.1.1　卡方检验的基本原理

检验的首要步骤是给出零假设：总体 X 服从某种分布，这里的样本认为是来自于总体 X。卡方检验的基本原理就是在给定零假设的条件下，根据样本值落在总体 X 的各个区间的频数与总体卡方落在该区间的期望频数的差值构造服从卡方分布的 Pearson 统计量。具体而言，总体 X 的取值被分成 k 个不相交的区间（子集），如果样本值落在总体的第 i 个区间的频数是 n_i，则当 k 足够大时，统计量

$$x^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

近似服从自由度为 k–1 的卡方分布，这里的 n 为样本总数，p_i 为总体落在第 i 个区间的概率值，称 np_i 为期望频数。从卡方统计量的计算公式可以看出，如果 x^2 值较大，则样本频数分布与期望频数分布差距较大；反之，如果 x^2 值较小，则样本频数分布与期望频数

分布较接近。SPSS 将自动计算 x^2 统计量的观测值，并依据卡方分布表计算观测值对应的概率 ρ 值是否大于显著性水平，不能拒绝零假设，便可以认为样本来自于服从某分布的总体。

7.1.2　卡方检验的实例

SPSS 的卡方分布检验的数据存放有两种方式：一种是定义一个变量存放所有的样本值，重复的样本值作为不同的个案保存；另一种是定义两个变量，一个用于存放不同的样本值，另一个存放该样本值的相应频数，但这时应将频数变量指定为加权变量。

【示例 7.1】卡方检验的实例。

例如，假设有 100 个摇奖数据，全部数据在表 7.1 中列出，检验摇奖数据是否均匀分布在 0~9。

表 7.1　100 个摇奖数据

3 2 2 8 5 4 2 8 0 4 9 4 8 7 7 4 2 2 1 6 4 7 2 9 8 4 6 7 6 2 0 7 5 9 7 4 4 4 7 5 6 8 1 7 2 0 3 3 4 8
0 7 1 1 6 0 7 2 2 1 2 5 8 1 7 8 6 4 2 8 6 4 9 6 0 4 3 4 3 8 1 7 2 8 4 2 0 1 8 3 7 3 7 8 6 1 3 6 4 3

SPSS 卡方检验的基本操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开卡方检验主对话框。

依次单击菜单“ 分析→非参数检验→卡方检验 ”命令，打开“ 卡方检验 ”对话框，如图 7.2 所示。



图 7.2 “ 卡方检验 ”对话框

03 检验变量的选择和设置。

在变量列表框中选择检验变量，单击 按钮，将其移动到“检验变量列表”列表框中。在本例中唯一的检验变量为“摇奖号码”。

“期望全距”选项组用于确定检验值的范围，其选项及含义如表 7.2 所示。

表 7.2 期望范围选项及含义

选　项	含　义
从数据中获取	默认选项。由数据的最小值和最大值确定的范围
使用指定的范围	用户指定范围。可在被激活的“下限”和“上限”文本框中输入检验范围的下限和上限。在所指定的范围内每个单独的数值产生一个分组子集，数值在指定范围外的个案将被剔除

“期望值”选项组用于指定期望频数 np_i 的方法，对应两个单选项，其含义如表 7.3 所示。

表 7.3 期望值选项及含义

选　项	含　义
所有类别相等	默认选项。所有子集的期望概率值都相同，即检验总体是否服从均匀分布
值	指定分组的期望概率值。在其后的文本框内可以输入相应的大于 0 的数值，输入的顺序与检验变量分组的递增顺序相对应，列表框中的第一个数值对应于检验变量的最小分组值，最后一个对应于最大分组值。可以单击“添加、更改、删除”按钮，以对输入的数值进行增加、更改和移除

04 “选项”设置。

在“卡方检验”对话框中，单击“选项”按钮，打开“卡方检验：选项”对话框，如图 7.3 所示。

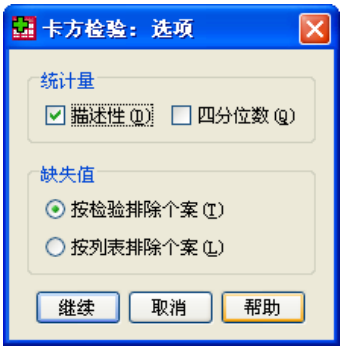


图 7.3 “卡方检验：选项”对话框

在该对话框的“统计量”选项组中包含选择统计量的两个复选项。其具体含义如表 7.4 所示。

表 7.4　统计量选项组中的选项及含义

选　　项	含　　义
描述性	描述性统计量，显示均值、标准差、最大值、最小值和非缺失个案数
四分位数	显示 4 分位数，即 25%、50%和 75%的百分位数

在该对话框的“ 缺失值 ”选项组中有对缺失值的处理方式，具体如表 7.5 所示。

表 7.5　缺失值选项组选项及含义

选　　项	含　　义
按检验排除个案	默认选项，用于在检验的个案中，将其中含有缺失值的个案剔除
按列表排除个案	选择该项，表示将含有缺失值的全部个案剔除

05 “ 精确检验 ”设置。

在“ 卡方检验 ”对话框中单击“ 精确 ”按钮，打开“ 精确检验 ”对话框，如图 7.4 所示。



图 7.4 “ 精确检验 ”对话框

“ 精确检验 ”对话框用于计算显著性水平 Sig.值的几种方法，具体含义如表 7.6 所示。

表 7.6　精确检验选项及含义

选　　项	含　　义
仅渐进法	默认计算显著性水平的方法。计算显著性水平是基于检验统计量的渐近分布假设。如果显著性水平小于 0.05，则检验结果被认为存在显著性差异。根据大数定理理论，渐近方法的显著性水平要求数据量足够大，如果数据量比较小，或者频数过低，则检验结果可能会失效
Monte Carlo（蒙特卡罗法）	蒙特卡罗估计法是精确显著性水平的无偏估计。它利用给定样本集通过模拟方法重复取样，计算显著性水平，该方法不要求渐近方法的假设。蒙特卡罗法是一个很有用的方法。对于处理不满足渐近假设的巨量数据，而无法得到精确的显著性水平时，可以选择该方法在“ 置信水平 ”文本框中输入计算显著性水平的置信度，在“ 样本数 ”文本框中输入取样数量
精确	精确计算显著性水平的方法可以得到精确的显著性水平，不需要渐近方法的假设，不足之处是计算量和所需内存太大。选择该项后，可以选择“ 每次检验的时间限制 ”复选项，即设置限制计算时间，默认时间限制是 5 分钟，超过该时间，系统将自动停止计算，给出计算信息

06 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TEST
    /CHISQUARE=摇奖号码
    /EXPECTED=EQUAL
    /MISSING ANALYSIS.
```

设置完毕后，单击“确定”按钮，执行卡方检验操作。

本例全部选择默认选项。

- 描述性统计量

表 7.7　描述性统计量

Descriptive Statistics

	N	Mean	Std.Deviation	Minimum	Maximum
摇奖号码	100	4.46	2.687	0	9

表 7.7 显示检验数据的描述性统计量，100 个数据均值为 4.46，标准差为 2.687，最小值分别为 0 和 9。

- 期望频数和观测频数

表 7.8　期望频数和观测频数

摇奖号码

	Observed N	Expected N	Residual
0	7	10.0	-3.0
1	9	10.0	-1.0
2	14	10.0	4.0
3	9	10.0	-1.0
4	16	10.0	6.0
5	16	10.0	-6.0
6	10	10.0	.0
7	14	10.0	4.0
8	13	10.0	3.0
9	4	10.0	-6.0
Total	100		

表 7.8 给出了各个分组的期望频数和观测频数以及两者之间的差值。

- 检验结果

表 7.9 给出的是单样来卡方检验结果。X^2=16.000，渐近方法的概率 ρ 值为 0.076>0.05，所以接受原假设，说明给定的 100 个摇奖号码的分布是均匀的。

表 7.9　单样本卡方检验结果

Test Statistics

	摇奖号码
Chi-Square[a]	16.000
df	9
Asymp.Sig.	.067

a．0 cells(.0%)have expected frequencies less than

5．the minimum expected cell frequency is 10.0.

注释：表 7.9 中，Chi-Square[a] 表示 X^2 统计量检验，df 表示样本的自由度，Asymp.Sig.表示渐近方法的概率 ρ 值。

7.2　二项分布检验

二项分布是统计学中常见的一种分布，具有非常重要的意义。

7.2.1　二项分布检验的基本原理

在实际问题中，有很多数据的取值只取两个值，例如，进行射击时只考虑击中和没击中两种情况；学生成绩只考虑及格和不及格两种情况；医学中某种诊断指标中的阴性和阳性等。通常在数据文件中将这两种情况用 0 和 1 表示。假设进行一次实验，出现 1 的概率为 ρ，那么进行 n 次相同的实验，则出现 1（或 0）的次数用随机变量表示，该随机变量服从参数为 ρ 的二项分布。

SPSS 提供的二项分布的检验过程，用于对给定样本数据检验其总体是否服从概率为指定数值的二项分布，其零假设 H_0：样本来自于参数为指定数值的二项分布。

对于小样本的二项分布检验，构造精确检验统计量；对于大样本的二项分布检验，利用大数定理构造近似服从正态分布的统计量。

如果检验统计量的概率 ρ 值小于显著性水平，则应拒绝零假设，认为样本不是来自于服从参数为给定值的二项分布总体；反之，如果检验统计量的概率 ρ 值大于显著性水平，则不能拒绝零假设，可以认为样本来自于服从给定参数的二项分布总体。

7.2.2　二项分布检验的实例

SPSS 中二项分布操作步骤如下。

01 打开二项分布对话框。

依次单击菜单“ 分析→非参数检验→二项式检验 ”命令，打开“ 二项式检验 ”对话框。如图 7.5 所示。

02 选择检验变量。

在对话框的左侧变量列表框中选择检验变量，单击 按钮，将其移动到右侧的“ 检

验变量列表”列表框，如图 7.5 所示。

对话框的“定义二分法”选项组给出定义两值的选项，即指定分类方法，有以下两个选项：从数据中获取和割点，其含义如表 7.10 所示。



图 7.5 “二项式检验”对话框

表 7.10 定义二分变量选项及含义

选　项	含　义
从数据中获取	默认选项，适用于指定的变量只有两个有效值且无缺失值的情况
割点	如果指定的变量超过两个值，则选择该项，并在其后的文本框内输入一个具体数值，比该值小的观测值为第一组，比该值大的观测值为第二组

在“检验比例”文本框中输入用于检验二项分布的概率值作为第一组变量的待检验概率值，默认概率值为 0.50，输入范围为 0.001~0.999。

03 选择和设置子对话框。

单击“选项”按钮，可以打开“二项检验：选项”对话框；单击“精确”按钮，可以打开“精确检验”对话框。这两个对话框的功能和设置方法和示例 7.1 相应的对话框完全相同，这里不再赘述。

04 执行操作。

单击“确定”按钮，执行二项分布检验操作。

【示例 7.2】抛硬币实验的二项分布检验。

某人做抛掷硬币实验，检验硬币正面出现的概率是否为 1/2。抛掷 60 次，出现正面记为 1，出现反面记为 0，记录的结果如表 7.11 所示。

表 7.11 60 个抛掷硬币数据

1 1 0 0 1 0 0 1 0 0 0 1 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 1 1 1 1 0 0 0 1 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 0 1 1

检验硬币正反面出现的次数是否服从概率为 0.5 的二项分布。

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开二项分布检验主对话框。

依次单击菜单“ 分析→非参数检验→二项式检验 ”命令，打开“ 二项式检验 ”对话框。将变量“ 硬币结果 ”移至“ 检验变量列表 ”列表框。

03 设置。

单击“ 选项 ”按钮，在“ 二项分布检验：选项 ”对话框中选择“ 描述 ”复选项。

04 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TEST
  /BINOMIAL (.50)= 硬币结果
  /STATISTICS  DESCRIPTIVES QUARTILES
  /MISSING ANALYSIS.
```

设置完毕后，单击“ 确定 ”按钮，执行卡方检验操作。

本例全部选择默认选项。

- 描述性统计量

表 7.12　二项分布检验的描述性统计量

Descriptive Statistics

	N	Mean	Td.Deviation	Minimum	Maximum	Percentiles		
						25th	Oht(Mediar)	75th
硬币结果	60	.48	.504	0	1	.00	.00	1.00

表 7.12 中给出了检验变量的描述性统计量，均值为 0.48，标准差为 0.504。

- 二项分布检验结果

表 7.13　二项分布检验的结果

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp.Sig.(2-tailed)
硬币结果	Group 1	1	29	.48	.50	.897[a]
	Group 2	0	31	.52		
	Total		60	1.00		

a．Based on Z Approximation.

表 7.13 中给出了二项分布检验的结果。从表中可以看出，第一组的频数为 29，第二组的频数为 31，检验的二项分布的概率值为 0.50，检验统计量为近似正态分布（大样本情况下），计算得到的概率 ρ 值为 0.897>0.05，所以接受零假设，可以认为硬币的正反面出现的概率相等，都为 1/2。

7.3 游程检验

游程检验（Runs Test）是一种利用游程的总数获得统计推断结论的方法。

7.3.1 游程检验的基本原理

首先，简要介绍以下游程检验的概念。

设 $x_1, x_2, ..., x_m$ 和 $y_1, y_2, ..., y_n$ 分别来自同一总体 X 和 Y 的随机样本数据，如果将这两个样本合并在一起，并按照从小到大的顺序排列，得到次序统计量：$z_1, z_2, ..., z_{m+n}$。

如果将次序统计量中来自于总体 X 的样本值记为 0，来自于总体 Y 的样本值记为 1，仍按原顺序表示次序统计量，则得到一个由 0 和 1 构成的序列。我们将连续出现 0 或者连续出现 1 的一组数分别称为 0 的游程或者 1 的游程，一个游程中包含 0 或者 1 的个数称为游程长度。

例如，如果随机抛掷硬币 20 次，观察正反面，得到序列 11100100100011011100，该序列中的 111、00、1、00 都是游程，其中有 0 的游程 5 个，长度为 1 的游程 3 个，长度为 2 的游程 4 个，长度为 3 的游程 3 个。

如果 0 和 1 是随机出现的，则出现连续的很多 0 或者连续的很多 1 的可能性很小，同时，1 和 0 的频繁交叉出现的可能性也很小，检验两组数据是否具有随机性，游程是其重要的依据。

在 SPSS 的游程检验过程中，利用游程构造检验统计量，判断样本数据的随机性，在大样本条件下，游程的抽样分布近似服从正态分布，SPSS 计算统计量，并根据正态分布的分布函数求出相应的概率 ρ 值，如果概率 ρ 值小于给定的显著性水平，则拒绝零假设，认为序列不具有随机性；反之，如果概率 ρ 值大于给定的显著性水平，则接受零假设，认为序列具有随机性。

利用游程检验方法既可以对次序统计量进行随机性检验，还可以对不同的两个总体进行显著性检验。

7.3.2 游程检验的实例

SPSS 中游程检验的具体操作步骤如下。

01 打开游程检验对话框。

依次单击菜单“分析→非参数检验→游程检验”命令，打开“游程检验”对话框。

02 指定检验变量。

在左侧的变量列表框中选择检验变量，单击 按钮，将其移动到右侧的“检验变量列表”列表框中。如图 7.6 所示。

03 确定游程数的割点。

在图 7.6 所示对话框的“ 割点 ”选项区域中确定划分两类的割点，其中小于割点的变量值构成第一类，其他个体构成第二类，割点复选项及含义如表 7.14 所示



图 7.6 “ 游程检验 ”对话框

表 7.14　割点复选项及含义

选　项	含　义
中位数	检验变量的中位数作为割点
众数	检验变量的众数作为割点
均值	检验变量的均值作为割点
设定	用户自定义割点。选择该项时，在右侧的文本框中输入自定义的割点值

04 选择和设置子对话框。

单击“ 选项 ”按钮，打开“ 执行检验：选项 ”对话框；单击“ 精确 ”按钮，可以打开“ 精确检验 ”对话框。这两个对话框的功能和设置方法和 7.1 节相应的对话框完全相同，这里不再赘述。

05 执行操作。

设置完毕后，在“ 游程检验 ”主对话框中，单击“ 确定 ”按钮，执行操作。

【示例 7.3】体育达标成绩的检验。

为了考察学生的体育达标成绩，从两组学生中分别选取了 15 和 14 名学生进行测试，测试成绩如表 7.15 所示。

表 7.15　两组学生体育测试成绩

组　别	得　分
甲组	67 67 70 70 73 74 78 81 83 84 84 87 87 97 97
乙组	63 66 71 74 74 75 76 77 77 78 79 82 86 87

检验两组学生的体育成绩测试是否有显著差别，其具体操作步骤如下。

01 建立数据文件。

将数据文件导入 SPSS 中。

02 对变量“分数”进行排序。

依次选择菜单“数据→排序个案”命令，进入“排序个案”对话框，如图 7.7 所示。在该对话框中，选择变量“分数”，将其移动到“排序依据”列表框；在“排列顺序”选项组中选择“升序”；单击“继续”按钮，返回到主对话框。

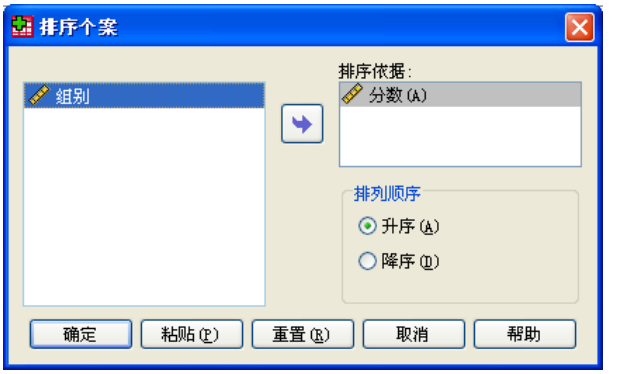


图 7.7 “排序个案”对话框

03 进行游程检验。

依次单击菜单“分析→非参数检验→游程检验”命令，打开“游程检验”对话框。将变量“组别”移动到“检验变量列表”列表框；在“割点”选项组中选择“平均值”作为割点；其他为默认选项。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
/*游程检验的操作命令语句
NPAR TESTS
    /RUNS(MEDIAN)=组别
    /MISSING ANALYSIS.
```

执行完“游程检验”后，得到游程检验的结果。

- 游程数

表 7.16 经过排序的分数

组　别	成　绩	组　别	成　绩
2	63	1	78
2	66	2	78
1	67	2	79
1	67	1	81
1	70	2	82
1	70	1	83
2	71	1	84
1	73	1	84
2	74	2	86

续表

组　别	成　绩	组　别	成　绩
2	74	1	87
1	74	2	87
2	75	1	87
2	76	1	97
2	77		

表 7.16 是对分数排序后的列表。从表中的组别列可以观察到经排序后的组别的游程数。

- 游程检验结果

表 7.17　游程检验结果

Runs Test

	组　别
Test Value[a]	1.48
Cases<Test Value	15
Cases>=Test Value	14
Total Cases	29
Number of Runs	16
Z	.007
Asymp.Sig.(2-tailed)	.995

a．Mean

注释：表 7.17 中，Test Value[a] 表示变量“ 组别 ”的均值，Cases<TestValue 和 Cases>=TestValue 分别表示小于 TestValue 和大于等于 TestValue 的个案，Total Cases 为总个案数，Number of Runs 为游程数，Z 表示 Z 检验统计量的值，Asymp.Sig.（2-tailed）为概率 ρ 值。

表 7.17 是游程检验的结果。检验值是变量“ 组别 ”的均值，为 1.48，该值将数据分成两组，其中在第一组有 15 个个案，在第二组有 14 个个案。游程数为 16 个，检验统计量的值为 0.007，因为概率 ρ 值为 0.995，远大于 0.05，所以接受零假设，即认为两组学生的体育成绩没有显著差异。

7.4　单个样本 K-S 检验

K-S 检验是柯尔莫哥洛夫-史密尔诺夫（Kolmogorov-Smirnov）检验的简称，是以俄罗斯科学家柯尔莫哥洛夫和史密尔诺夫的名字命名的一种非参数检验方法。该方法是一种根据样本数据推断其来自的总体是否服从某一特定理论分布，也是一种拟合优度检验方法。

7.4.1　单个样本 K-S 检验的基本原理

由于卡方拟合优度检验需要将样本空间分成不相交的子集，所以存在较多主观因素，

分组方法不同有时会导致检验结论不同或者检验方法失效的情况；而 K-S 检验方法在一定程度上克服了卡方检验的缺点，是比卡方检验更精确的一种非参数检验方法。

K-S 检验是将样本的累积分布函数与确定的理论分布函数相比较，根据差值的大小确定样本是否来自于给定的总体。K-S 检验按照绝对值计算两个分布函数之间的最大差异构造检验统计量，其零假设 H_0：样本来自于给定分布的总体。

在大样本条件下，检验统计量近似服从正态分布，SPSS 自动计算检验统计量的概率 ρ 值，并根据概率 ρ 值是否大于显著性水平（一般取 0.05），决定是否接受或拒绝零假设。

K-S 检验方法可以对正态分布、均匀分布、泊松分布和指数分布进行检验，其检验参数可以根据给定的样本进行估计。

例如，对于正态分布，将样本均值和样本标准差作为理论分布的参数；对于均匀分布，将样本的最小值和最大值作为其取值范围；对于泊松分布和指数分布，将样本均值作为其理论参数。

7.4.2 单个样本 K-S 检验的实例

SPSS 中单样本 K-S 检验的操作步骤如下。

01 打开单样本 K-S 检验对话框。

依次单击菜单“分析→非参数检验→1-样本 K-S 检验”命令，进入“单样本 Kolmogorou-Smirnou 检验”（简称“单样本 K-S 检验”）对话框，如图 7.8 所示。

02 指定检验变量。

在“单样本 K-S 检验”对话框左侧的变量列表框中选择检验变量，单击 按钮，将其选入到右侧的“检验变量列表”列表框。



图 7.8 “单样本 K-S 检验”对话框

03 确定检验的分布类型。

在图 7.8 所示的对话框中，“检验分布”选项组用于指定检验的分布类型，SPSS 提供了 4 种常用的分布，有：常规（默认选项）、相等、泊松和指数分布。

04 选择和设置子对话框。

单击图 7.8 所示对话框中“ 选项 ”按钮，可以打开“ 单样本 K-S 检验：选项 ”对话框；单击“ 精确 ”按钮，可以打开“ 精确检验 ”对话框。这两个对话框的功能和设置方法请参考 7.1 节相应的对话框。

05 执行操作。

在“ 单样本 K-S 检验 ”主对话框中，单击“ 确定 ”按钮，执行操作。

【示例 7.4】车祸事故数的泊松分布检验。

保险公司的车险分析员对 500 例驾驶员的车祸事故进行分析，随机抽取 500 例样本数据，包括性别和近五年的事故记录的资料（如表 7.18 所示），分析员利用 K-S 检验方法确定事故数是否服从泊松分布。

表 7.18　500 例驾驶员近 5 年的事故数

性别	事 故 数
男	2 1 1 0 1 1 1 3 5 1 3 3 3 5 0 2 1 2 3 3 1 4 3 0 1 3 0 6 4 3 2 1 1 0 3 1 1 1 1 0 2 1 0 2 2 1 6 1 2 3 2 1 2 2 2 0 0 2 4 4 1 2 2 1 4 3 2 2 4 0 4 1 0 2 2 2 0 2 1 2 0 2 3 4 4 2 0 1 1 1 3 1 3 5 3 0 4 1 1 1 0 1 0 0 4 0 0 5 2 5 4 0 1 2 5 3 0 2 0 4 2 4 0 3 3 1 2 4 3 1 3 0 3 5 2 0 3 2 3 1 0 3 0 0 2 4 1 1 5 0 1 0 1 0 4 4 0 1 2 1 0 3 2 6 1 7 1 2 1 1 0 1 2 1 3 0 1 2 4 0 3 4 0 1 6 3 0 6 1 0 1 1 0 2 3 2 2 3 3 1 3 0 1 2 2 1 3 0 4 1 1 1 5 3 2 0 1 2 5 1 1 2 1 3 1 2 3 2 2 1 0 3 2 1 1 1 4 2 4 1 1 2 6 7 0 1 2 5 1 6
女	1 3 1 0 0 0 2 5 5 1 0 1 3 4 0 0 2 1 2 2 3 3 2 2 0 5 1 1 1 2 0 1 1 3 0 1 0 2 3 1 0 6 0 0 4 6 1 1 1 1 0 4 2 1 0 2 1 1 2 0 2 1 1 1 2 1 0 2 0 2 3 0 3 1 1 4 2 0 1 2 2 5 2 3 1 2 1 1 5 0 0 0 1 2 1 1 0 0 2 4 0 5 0 1 0 0 4 0 5 2 2 0 1 0 4 1 2 3 0 0 2 0 0 1 0 0 4 3 2 1 1 3 1 2 3 0 2 0 4 1 5 0 3 2 2 0 0 1 1 0 1 0 1 4 1 2 2 1 0 1 2 2 1 2 2 0 1 1 2 2 2 2 1 1 1 1 3 0 0 3 1 1 0 1 0 0 2 1 0 3 1 1 3 2 3 0 0 1 2 0 2 1 0 0 1 1 3 0 0 0 2 0 0 0 0 0 1 2 3 1 3 4 4 0 0 2 3 0 0 1 3 3 0 1 4 2 4 1 4 2 4 1 0 0 2 0 1 0 2

该问题的具体操作步骤如下。

01 建立数据文件。

将数据导入 SPSS 中。

02 选择检验对话框。

依次单击菜单“ 分析→非参数检验→单样本 K-S 检验 ”命令，进入“ 单样本 K-S 检验 ”对话框。选择变量“ Number of accident past 5 years[accident] ”将其移动到“ 检验变量列表 ”列表框中；在“ 检验分布 ”选项组中选择“ 泊松 ”。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TESTS
   /K-S(POISSON)= accident
   /STATISTICS DESCRIPTIVES
   /MISSING ANALYSIS.
```

04 单击“ 确定 ”按钮，执行操作。

执行完“ 单样本 K-S 检验 ”后，得到单样本 K-S 检验的结果。

表 7.19 为单样本 K-S 检验结果，从中可以看出：

- 检验的数据是基于 500 名司机的事故数，第二行中的 N 表示检验的个案数。

- Poisson Parameter 表示泊松分布的参数值，泊松分布利用样本均值表示其唯一的参数，样本均值为 1.72，即 5 年内每个司机的平均事故数。

表 7.19　单样本 K-S 检验结果

One-Sample Kolmogorov-Smirnov Test

		Number of accidents past 5 years
N		500
Poisson Parameter[a,b]	Mean	1.72
Most Extreme Differences	Absolute	.065
	Positive	.065
	Negative	-.041
Kolmogorov-Smirnov Z		1.460
Asymp.Sig.(2-tailed)		.028

a．Test distribution is Poisson.

b．Calculated from data.

注释：表 7.19 中，*N* 表示检验的个案数，Poisson Parameter[a,b] 表示泊松分布的参数值（Mean 为泊松分布的样本均值），MMost Extreme Differences 表示样本数据和理论数据的最大差值（Absolute 为绝对差值，Positive 为正差值，Negative 为负差值），Kolmogorov-Smirnov Z 为 K-S 正态检验统计量的值，Aaymp.Sig.(2 tailed)为双尾渐近概率 ρ 值。

- Most Extreme Differences 表示样本数据和理论数据的最大差值，最大绝对差值为 0.065，最大正差值为 0.065，最大负差值为-0.041，最大（正、负）差值的大小可以判断理论分布和经验分布之间的差距，为计算检验统计量提供了直观的分析。
- Kolmogorov-Smirnov Z 为 K-S 正态检验统计量的值，这里为 1.460。
- Asymp.Sig.(2 tailed)为双尾渐近概率 ρ 值，为 0.028，小于给定的显著性水平 0.05，所以零假设被拒绝，即认为给定的 500 例司机的事故数据不能认为来自于参数为 1.72 的泊松分布。

考虑性别因素，下面对实例进一步分析。

在示例 7.4 中 K-S 检验假设被拒绝，原因有两个：选择不合适的理论分布或者不正确的分布参数。根据泊松分布的性质，仍然有理由认为事故数分布为泊松分布。重新考察上述检验过程，会发现检验过程中没有考虑性别因素，500 例数据中男女性别数相同，这不符合实际情况。

下面考虑不同性别的 K-S 检验，所以先对 500 例数据按照性别变量进行拆分。步骤如下。

01 依次单击菜单“数据→拆分文件”命令，打开“拆分文件”对话框，如图 7.9 所示。在其中选择“比较组”，然后将变量“Sex of insured[gender]”移动到“分组方式”列表框，单击“确定”按钮，确认对性别的分组。

02 对变量“ Number of accident past 5 years[accident] ”进行单样本 K-S 泊松检验操作，选择“ 选项 ”中的“ 描述 ”选项，其他操作和前面相同。



图 7.9　对数据按照性别进行拆分

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
/*拆分数据文件操作
NPAR TESTS
    /K-S(POISSON)= accident
    /STATISTICS DESCRIPTIVES
    /MISSING ANALYSIS.
    SORT CASES BY gender .
    SPLIT FILE
    LAYERED BY gender .

/*K-S 检验的命令语句
NPAR TESTS
    /K-S(POISSON)= accident
    /STATISTICS DESCRIPTIVES
    /MISSING ANALYSIS.
```

04 单击“ 确定 ”按钮，执行操作。

执行完“ 单样本 K-S 检验 ”后，得到单样本 K-S 检验的结果。

- 描述性统计量

表 7.20　描述性统计量

Descriptive Statistics

Sex of insured		*N*	Mean	Std.Deviation	Minimum	Maximum
Male	Number of accidents Past 5 years	250	1.98	1.608	0	7
Female	Number of accidents Past 5 years	250	1.47	1.412	0	6

注释：表 7.20 中，Male 和 Female 分别表示男性司机和女性司机。*N* 为过去 5 年发生事故的数量，Mean 为平均事故数，Std.Deviation 为标准差偏差指数，Minimum 和 Maximum 为样本统计量的最小值和最大值。

表 7.20 给出了描述性统计量的列表。从表中的描述性统计量可以看出，男性的平均事故数增加到 1.98，而女性的平均事故数则减少的 1.47。

- 拆分后的检验结果

表 7.21　拆分后的单样本 K-S 检验

One-Sample Kolmogorov-Smirnov Test

Sex of insured			Number of accidents past 5 years
Male	*N*		250
	Poisson Parameter[a,b]	Mean	1.98
	Most Extreme Differences	Absolute	.047
		Positive	.047
		Negative	-.033
	Kolmogorov-Smirnov Z		.750
	Asymp.Sig.(2-tailed)		.627
Female	*N*		250
	Poisson Parameter[a,b]	Mean	1.47
	Most Extreme Differences	Absolute	.074
		Positive	.074
		Negative	-.042
	Kolmogorov-Smirnov Z		1.164
	Asymp.Sig.(2-tailed)		.133

a．Test distribution is Poisson.

b．Calculated from data.

注释：表 7.21 中，Number of accidents past 5 years 表示过去 5 年发生的事故数，Poisson Parametera,b 表示泊松分布的参数值（Mean 为泊松分布的样本均值），Most Extreme Differences 表示样本数据和理论数据的最大差值（Absolute 为绝对差值，Positive 为正差值，Negative 为负差值），Kolmogorov-Smirnov Z 为 K-S 正态检验统计量的值，Aaymp.Sig.(2 tailed)为双尾渐近概率 ρ 值。

表 7.21 显示的是经过对性别拆分以后，分别对男性司机和女性司机进行 K-S 泊松检验后的结果。

从结果中可以看出，当男性司机的事故均值从 1.72 提高到 1.98 以后，以此作为泊松分布检验的参数，得到的概率 ρ 值提高到 0.627，大于 0.05，通过显著性检验，可以认为男性司机的事故数服从泊松分布；对女性司机也得到了同样的结论。

所以，利用单样本 K-S 检验，不仅需要选择合适的理论分布，还需要精确的参数估计值。

7.5　两个独立样本的非参数检验

在前面章节中我们讨论了两个独立样本的参数检验问题，利用 T 检验可以解决两个总

体是否存在显著性差异问题，并且该检验方法具有计算简单、稳健性强等特点；但 T 检验方法要求两个独立样本来自的总体服从正态分布，甚至是对两个总体的方差也要求相等或者已知。在实际问题中，有很多时候对给定样本的分布形态并不是很了解，更不用说总体的参数情况，这时，可以利用两个独立样本的非参数检验过程对两个总体的分布是否存在显著性差异进行检验。

两个独立样本的非参数检验是通过对两个独立样本的分析，来推断来自两个总体的分布是否存在显著性差异，之所以称为非参数检验，是因为检验过程不需要已知总体的分布，也不需要已知总体的参数，甚至样本的数值可以不是定距型变量值。

SPSS 提供了多种两个独立样本的非参数检验方法，有 Mann-Whitney U 检验、K-S 检验、W-W 游程检验和极端反应检验。

下面以两个独立样本 Mann-Whitney U 检验来说明两个独立样本检验的具体操作步骤。

7.5.1　两个独立样本 Mann-Whitney U 检验

两个独立样本 Mann-Whitney U 检验可用于检验两个独立样本是否来自于同一总体，它是最常用的两个独立样本的非参数检验方法。该检验法的零假设是 H_0：两个独立样本来自于无显著差异的总体。Mann-Whitney U 检验是利用两个独立样本的秩的比较得到的。

首先，将两样本数据按照升序合并在一起，计算第一组样本的观察值大于第二组样本观察值的次数；然后，再计算第二组样本的观察值大于第一组样本的观察值的次数，分别利用 U1 和 U2 表示。如果 U1 和 U2 比较接近，则说明两个独立样本有可能来自于同一个整体；当然，两样本的差异不仅和 U1、U2 有关，还和两样本的容量有关。

利用 U1 和 U2 构造检验 Mann-Whitney U 统计量，样本较小时，U 统计量服从 Mann-Whitney 分布；当样本容量较大时，则近似服从正态分布，利用分布函数表，SPSS 将自动计算出 U 统计量的值和概率 ρ 值。

如果概率 ρ 值小于给定的显著性水平，则拒绝零假设，可以认为两个独立样本来自的总体存在显著差异；反之，如果概率 ρ 值大于给定的显著性水平，则接受零假设，两个独立样本来自的总体的分布无显著性差异。

由于 Mann-WhitneyU 检验实际上近似于一种秩和检验，检验过程的统计量和样本值的大小没有关系，只和排序有关，所以 Mann-Whitney U 检验不仅可以用于样本值为连续型数据，而且还可以用于定序型数据。

7.5.2　两个独立样本非参数检验的实例

两个独立样本的非参数检验的 SPSS 操作具体步骤如下。

01 选择对话框。

依次单击菜单“分析→非参数检验→两个独立样本检验”命令，打开“两个独立样本检验”对话框，如图 7.10 所示。



图 7.10 “ 两个独立样本检验 ” 对话框

02 指定检验变量和分组变量。

在变量列表中选择检验变量，将其移动到右侧的“ 检验变量列表 ” 列表框；在变量列表中选择分组变量，将其移动到右侧的“ 分组变量 ”列表框。选择分组变量以后，单击“ 定义组 ” 按钮，打开如图 7.11 所示的“ 两独立样本：定义组 ” 对话框，在该对话框确定分组值，分别在“ 组 1 ” 和“ 组 2 ” 文本框中指定分组变量的值。

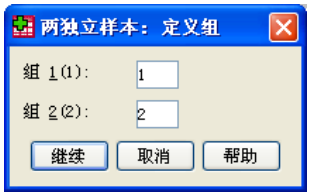


图 7.11 “ 两独立样本：定义组 ” 对话框

03 确定检验方法。

在“ 两个独立样本检验 ” 对话框“ 检验类型 ” 选项组中选择用来进行检验的方法，有 4 个复选项，如表 7.22 所示。

表 7.22 检验类型

检验类型	含　义
Mann-Whitney U	即 Wilcoxon 秩和检验，是最常用的两个独立样本检验。可以检验两样本所在的总体是否具有相同分布，如果两样本来自同一总体，则在两样本上等级的分布应该是比较随机的。因此 U 若出现极端值表示非随机；当样本例数小于 30 时，则计算双侧精度概率 ρ 值。对于大样本，则 U 被转换成正态分布的 Z 统计量
Kolmogorov-Smirnov Z	Kolmogorvo-Smirnov 两独立样本 Z 检验，是检验两个分布的位置和形状分布差异的方法。该方法主要用于基于两个样本的秩的累积分布函数的最大绝对差异；当差异有显著性时，可认为两个分布不同

续表

检验类型	含　义
Moses 极限反应	Moses 极限反应检验。由于观察值的范围很容易受到极端值的影响，而使计算结果位置偏离，以致无法真正显示两个样本的观测值是否有差异。尽量控制此问题所采用的措施是将某些计算观察值删除
Wald-Wolfowitz 游程	Wald-Wolfowitz 游程检验。用于检验两样本是否来自相同的总体。在计算游程数中，如果两样本间发生的数值相同，则尝试不同的排列顺序组合，以求得最大和最小的游程数；当样本容量小于或等于 30 时，计算单侧精确 ρ 值

【示例 7.5】抑郁症患者治疗效果的检验。

有两种方法治疗抑郁症，一种是药物治疗方法，另一种是药物和精神结合的综合治疗方法。有 13 位病人采用第一种治疗方法，17 位病人采用第二种治疗方法，一个疗程以后比较两组病人的几个特征是否有显著不同。比较项目为精神、体力和饮食三个方面，病人针对每一个项目的变化分别为从 0 到 4 的 5 种情况，0 代表最差，4 代表最好。表 7.23 为病人统计数据，利用 Mann-Whitney U 检验法检验两组病人的 3 个项目有无显著差异。

表 7.23　两种不同治疗方法下的病人情况分布

	第一组（1）	第二组（2）
精神	0434322211421	1343344340424344 4
体力	1212002322433	2434313344322233 3
饮食	4322402222023	3343322224122020 2

该问题的具体操作步骤如下。

01 建立数据文件。

将数据导入 SPSS 中。

02 选择检验对话框。

依次单击菜单“分析→非参数检验→两个独立样本检验”命令，进入“两个独立样本检验”对话框。

03 确定检验变量和分组变量。

选择变量“精神、体力、饮食”将其移动到“检验变量列表”列表框中，作为检验变量；将变量“组别”移动到“分组变量”列表框，指定为分组变量，并在“两个独立样本检验：定义分组”对话框“组 1”和“组 2”文本框输入分组变量的值 1 和 2。

04 确定检验方法。

在“两个独立样本检验”对话框的“检验类型”选项组中选择“Mann-Whitney U”检验法。

05 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TESTS
  /M-W= 精神 体力 饮食   BY 组别(1 2)
  /MISSING ANALYSIS.
```

06 单击“确定”按钮，执行操作。

执行完“两个独立样本检验”后，得到检验结果。

- 平均秩次

表 7.24　两种不同治疗方法下病人的 3 个项目的 Mann-Whitney U 检验秩次

Ranks

组　别		N	Mean Rank	Sum of Ranks
精神	药物治疗	13	11.88	154.50
	综合治疗	17	18.26	310.50
	Total	30		
体力	药物治疗	13	11．46	149．00
	综合治疗	17	18．59	316．00
	Total	30		
饮食	药物治疗	13	15．38	200．00
	综合治疗	17	15．59	265．00
	Total	30		

表 7.24 给出了两组不同项目的秩的总和以及平均秩的数值。从表中可以看出，在精神项目方面，第一组的秩和为 154.50，平均秩为 11.88；第二组的秩和为 310.50，平均秩为 18.26。现在，在精神项目方面，第二组的平均秩比第一组的平均秩大得多，差距较大；类似地，在体力和饮食项目中，两组之间的平均秩相差仍很大，但在饮食项目中，两组之间却相差很小。

- 检验结果

表 7.25　两种不同治疗方法下病人的 3 个项目的 Mann-Whitney U 检验结果

Test Statistics[b]

	精　神	体　力	饮　食
Mann-Whitney U	63.500	58.000	109.000
Wilcoxon W	154.500	149.000	200.000
Z	-2.054	-2.292	-.068
Asymp.Sig.(2-tailed)	.040	.022	.946
Exact Sig. [2*(1-tailed Sig.)]	.048[a]	.028[a]	.967[a]

a．Not corrected for ties.

b．Grouping Variable:组别

表 7.25 表示两种不同治疗方法下病人的三个项目的 Mann-Whitney U 检验结果。从表中可以看出，在精神项目方面，两组之间的 Mann-Whitney U 统计量值为 63.500，Wilcoxon W 统计量的值为 154.500，正态统计量（Z）的值为-2.054。由于样本容量是小样本，因此可以采用 U 统计量的精确概率 ρ 值为 0.048，同时也给出了利用渐近方法计算的概率 ρ 值为.040；两种方法的概率 ρ 值都小于 0.05，拒绝零假设。所以，从精神变化来看，两种治

疗方法的效果存在显著差异；类似地，在体力方面，两种方法也存在显著差异；从饮食方面看，由于精确概率 ρ 值为 0.967，大于 0.05，两种治疗方法没有显著差异。

7.6　K 个独立样本的非参数检验

如果需要检验差异的独立样本不仅仅是两个，而是有三个或更多的时候，SPSS 提供的非参数检验方法包括 Cruskal-Wallis H（克-瓦氏单向方差分析）、Median（中数检验法）和 Jonckheere-Terpstra 方法。

适用于多个独立样本差异性检验的 SPSS 数据包含一个分组变量（自变量）、一个检验变量（因变量）。分组变量把试验区分为多个组，检验变量要求至少是顺序变量，当然也可以是等距或等比变量。

Cruskal-Wallis（克-瓦氏单向方差分析）直接对应于参数检验中的一维方差分析（One-Way ANOVA），通常用来检验几个样本是否来自于同一个总体。

Median（中数检验法）事实上是列联表分析，SPSS 首先把每个变量中大于中位数的观测值分为一组，小于或等于中位数的分为另一组，然后计算 x^2 值，精确度较低。

Jonckheere-Terpstra 方法（J-T 方法）与克-瓦氏单向方差分析近似，当分组变量为等级变量时，其精度比克-瓦氏单向方差分析要高。

如果检验结果显示存在显著差异，则各个独立样本必须做进一步的两两比较，例如采用 Mann-Whitney U 检验，以确定差异所在。

克-瓦氏单向方差分析和中位数检验法都有一定的前提假设，具体如下。

（1）克-瓦氏单向方差分析的前提假设

- 分布相同

检验变量中的各组数据总体呈连续分布，而且属于同类分布。如果任意两总体的得分不同，SPSS 只能检验两总体是否属于同一类型的分布，而不能判断两组数据的中位数是否相同。

- 随机性与独立性

试验数据来自随机抽样，它们之间的得分相互独立。

- 样本量

检验结果中的 x^2 值仅仅是估计值，样本越大结果越精确。样本量等于或超过 30 时，根据 x^2 值计算的显著性水平 ρ 值会相当精确。

（2）中位数检验法的前提假设

- 独立性

试验数据之间是相互独立的。

- 样本量

当样本量足够大时，中位数检验法所使用的列联表才会生成基于近似 x^2 分布的统计量，检验结果才会比较精确。

【示例 7.6】多个独立样本的非参数检验实例。

大脑损伤有时会导致语言功能的丧失或受损，成为失语症患者，而且大脑的不同部位损伤可能导致语言功能丧失的程度不同。一项研究比较了不同大脑部位损伤的患者的语言受损程度，其中 9 位额叶损伤患者，8 位颞叶损伤患者和 6 位顶叶损伤患者。研究采用了词图匹配的方法，即同时给被试人员呈现两类卡片，一类卡片是一些物体的图片，另一类是这些物体的名称，要求被试人员把这些卡片一一匹配起来。每正确匹配一对卡片得一分。三类患者的得分如表 7.26 所示。

表 7.26　三类患者测试得分表

患者类别	得　分								
额叶损伤患者	16	15	16	26	17	14	15	28	27
颞叶损伤患者	13	16	12	9	14	16	10	15	
顶叶损伤患者	27	15	24	23	27	30			

试问这三类患者在词图匹配中的得分有没有差异。

解决该问题的具体操作步骤如下。

01 建立数据文件。

将数据导入 SPSS 中，如图 7.12 所示。

通过分析发现，该数据包含两个变量，一个为分组变量（自变量），即“ 组别 ”，有三个水平，1 代表额叶损伤、2 代表颞叶损伤、3 代表顶叶损伤；另一个为检验变量（因变量），即“ 得分 ”。

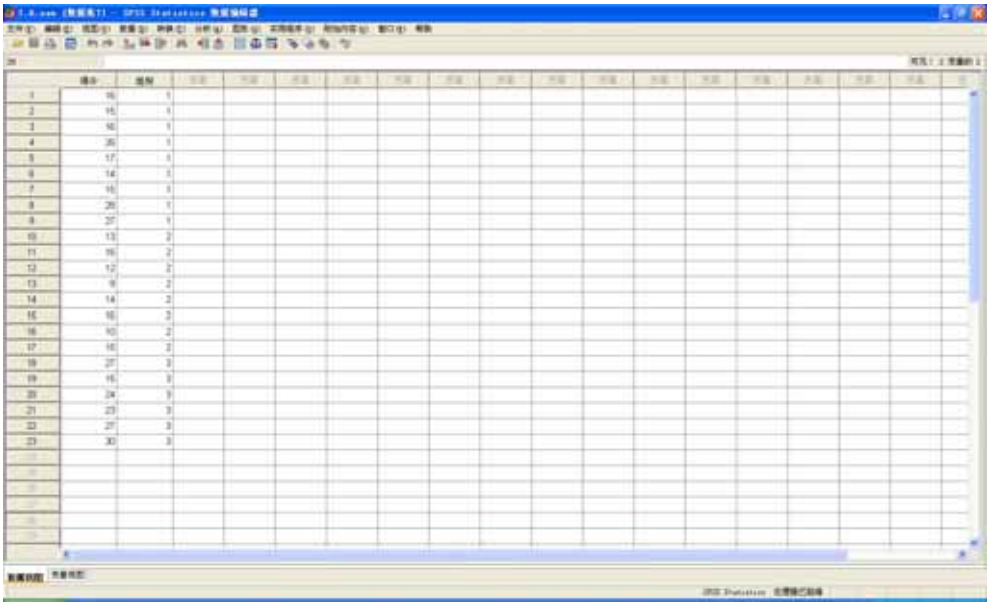


图 7.12　将数据文件导入 SPSS

02 打开检验对话框。

依次单击菜单“ 分析→非参数检验→K 个独立样本检验 ”命令，进入“ 多独立样本检验 ”对话框，如图 7.13 所示。



图 7.13 “ 多个独立样本检验 ”对话框

03 确定检验变量和分组变量。

选择变量“ 得分 ”将其移动到“ 检验变量列表 ”列表框中，作为检验变量；将变量“ 组别 ”移动到“ 分组变量 ”列表框，指定为分组变量，如图 7.13 所示。在“ 多自变量样本：定义范围 ”对话框“ 最小值 ”和“ 最大值 ”文本框输入分组变量的值：1 和 3，如图 7.14 所示。

04 确定检验方法。

在“ 多个独立样本检验 ”对话框的“ 检验类型 ”选项组中选择“ Kruskal-Wallis H ”检验法。

05 确定显著性水平计算方法。

在图 7.13 所示对话框中，单击“ 精确 ”按钮，打开“ 精确检验 ”对话框，选择“ 精确 ”方法，默认时间限制在 5 分钟以内，如图 7.15 所示。

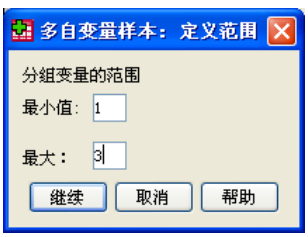


图 7.14 “ 多个独立样本：定义范围 ”对话框



图 7.15 “ 精确检验 ”对话框

06 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TESTS
   /K-W=得分   BY 组别(1 3)
   /MISSING ANALYSIS
   /METHOD=EXACT TIMER(5).
```

07 单击“确定”按钮，执行操作。

执行完“多个独立样本检验”后，得到检验结果。

- 描述性统计量

表 7.27　各组样本量及等级均值

Ranks

组　别		N	Mean Rank
得分	额叶损伤	9	13.61
	颞叶损伤	8	6.13
	顶叶损伤	6	17.42
	Total	23	

表 7.27 为各组样本量及等级均值，从表中可以看出，额叶损伤有 9 例，均值为 13.61，颞叶损伤有 8 例，均值为 6.13，顶叶损伤有 6 例，均值为 17.42。

- 统计检验结果

表 7.28　统计检验结果

Test Statistics[a,b]

	得　分
Chi-Square	10.467
df	2
Asymp.Sig.	.005
Exact Sig.	.002
Point Probability	.000

a．Kruskal Wallis Test

b．Grouping Variable:组别

注释：表 7.28 中，Chi-Squarea 表示 X^2 统计量检验，df 表示样本的自由度，Asymp.Sig.表示渐近方法的概率 ρ 值，Exact Sig.表示精确方法计算得到的 ρ 值，Point Probability 表示样本统计量的点概率。

表 7.28 为统计检验结果。克-瓦氏单向方差分析显示三类不同脑损伤患者的识字成绩存在显著差异，X^2（2，N=23）=10.467，精确方法计算的 ρ 值为 0.002，远小于 0.05，故不接受检验，即三类不同脑损伤患者的识字成绩是存在显著差异的。

- 两两比较

尽管我们发现统计检验的效果显著，但无法判断具体差异存在于哪两组之间，必须通过进一步检验来确定。我们采用 Mann-Whitney U 检验对三组被试人员的成绩进行两两比较。

通过主菜单“ 分析→非参数检验→两个独立样本检验 ”命令打开对话框，如图 7.16 所示。确定检验方法为“ Mann-Whitney U ”检验。



图 7.16 “ 两个独立样本检验 ”对话框

单击图 7.16 所示对话框“ 定义组 ”按钮，打开如图 7.17 所示对话框，在“ 组 1 ”和“ 组 2 ”文本框中分别要求比较组 1 和组 2、组 1 和组 3、组 2 和组 3。在结果输出中，SPSS 将分别输出相应的检验结果。

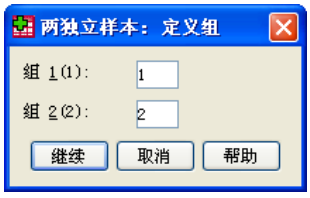


图 7.17 “ 两独立样本：定义组 ”对话框

单击图 7.16 所示对话框“ 精确 ”按钮，打开如图 7.18 所示“ 精确检验 ”对话框，选择“ 精确 ”检验法，默认时间限制在 5 分钟以内。



图 7.18 “ 精确检验 ”对话框

我们仅列出第 1 组和第 2 组进行比较的检验结果，如表 7.29 所示。

表 7.29　组 1 和组 2 比较检验结果

Test Statistics[b]

	得　分
Mann-Whitney U	10．500
Wilcoxon W	46．500
Z	-2.477
Asymp.Sig.(2-tailed)	.013
Exact Sig. [2*(1-tailed Sig.)]	.011[a]

a．Not corrected for ties.

b．Grouping Variable:组别

注释：表 7.29 中，Mann-Whitney U、Wilcoxon W 和 Z 表示三种不同的检验统计量，Asymp.Sig.表示渐近方法的概率 ρ 值，Exact Sig.表示精确方法计算得到的 ρ 值。

从表 7.29 可以看出，Mann-Whitney U 检验结果表明，额叶患者与颞叶患者的识字成绩存在显著差异。

7.7　两个相关样本的非参数检验

样本相关只有两个时，通常属于两种设计：一种是重复测量设计，即同一被试对象接受不同实验条件下的测试，目的在于检验被试对象在不同实验条件下的变化是否显著；另一种是配对样本设计，即被试对象经过配对后，分别接受不同的实验处理，目的在于检验两组被试对象在两种实验条件下是否差异显著。

在参数检验方法中，配对样本的 T 检验适用于上述两种设计。而相应的参数检验方法包括四种：Wilcoxon、Sign、McNemr 和 Marginal Homogeneity。

Wilcoxon 又称符号等级检验法，应用最广泛。Sign 又称符号检验法，统计精确度略低一些。McNemr 只适用于二分的相关变量。Marginal Homogeneity 是 McNemr 方法的扩展，可以检验多重反应的变量，但只限于顺序变量。后两种方法非常适用于前测—后测的实验设计。

两种设计下的 SPSS 数据文件都要求每个观测对象有成对的得分，即在两个变量上都有各自的得分。其中 Wilcoxon 与 Sign 检验两变量的中位数是否有显著差异。McNemr 方法适用的变量只包含两个分类（两个变量值），它检验被试对象同时属于两变量同一分列的概率是否有显著差异。

两个相关样本的非参数检验需要满足以下前提假设条件。

- 随机性与独立性

配对样本设计中的每对被试对象都来自于总体的随机抽样，各对被试对象之间保持相

互独立。

- 样本量

样本量越大，SPSS 输出的近似 Z 检验结果就越精确，当样本量达到 26 或更大时，检验结果会相当精确。

- 连续性与对称性（只适用于 Wilcoxon 方法）

两变量的差值总体呈连续分布，有对称性。

【示例 7.7】两个相关样本的非参数检验实例。

在示例 7.6 中，没有对老年痴呆症患者和脑中风患者的有关变量进行严格的匹配。在一项跟进的研究中，研究者精心挑选了被试对象，在年龄、性别、职业、受教育程度、患病时间等方面对老年痴呆症患者（A 组）和脑中风患者（B 组）进行了一一匹配，共挑选出 7 对被试对象（7 个老年痴呆症患者和 7 个脑中风患者）。研究者对他们的汉字识别情况进行了考察，从汉语常用字库中随机抽取了 30 多个汉字进行测试，要求患者逐次读出这些汉字。正确读出一个汉字计 1 分，读错不得分，也不扣分。两组患者的得分如表 7.30 所示。

表 7.30　两组患者测试得分表

患者类别	得　分						
A 组	7	3	9	19	21	5	8
B 组	2	6	4	17	22	4	2

试分析老年痴呆症患者和脑中风患者的识字成绩有没有差异。

该问题的具体操作步骤如下。

01 建立数据文件。

将数据导入 SPSS 中，如图 7.19 示。

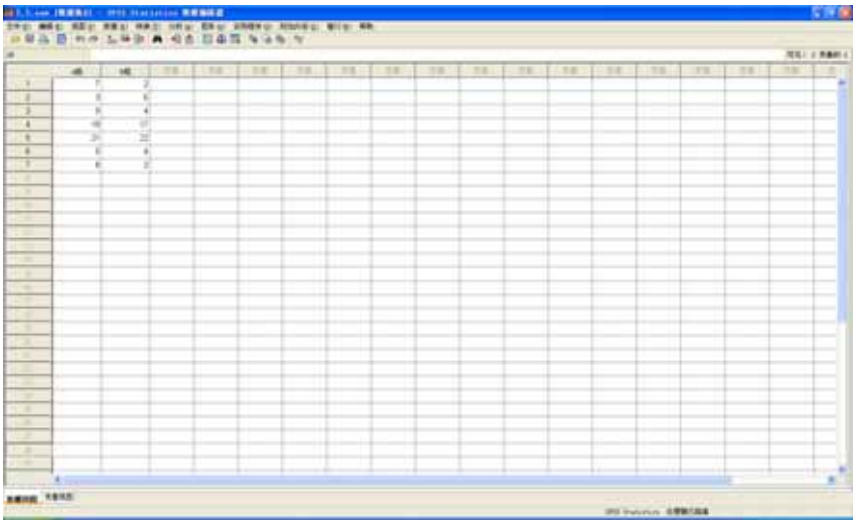


图 7.19　将数据文件导入 SPSS

通过分析发现，该数据包含两个变量，“ A 组 ”和“ B 组 ”分别代表老年痴呆症患者和脑中风患者的识字成绩，两样本容量都是 7。

02 打开检验对话框。

依次单击菜单“ 分析→非参数检验→两个关联样本 ”命令，进入“ 两个关联样本检验 ”对话框，如图 7.20 所示。



图 7.20 “ 两个关联样本检验 ”对话框

03 确定检验变量。

选中一个变量，按下“ Ctrl ”键再单击另一个目标变量将其选中，单击 按钮，将其移动到“ 检验配对列表 ”列表框中，作为检验变量。

04 确定检验方法。

在“ 两个相关样本检验 ”对话框的“ 检验类型 ”选项组中选择“ Wilcoxon ”检验法。

05 确定显著性水平计算方法。

在图 7.20 所示对话框中，单击“ 精确 ”按钮，打开“ 精确检验 ”对话框，如图 7.21 所示。对话框显示有三种可选方法。由于实际中使用的样本量比较小，每个样本只包含 7 位被试对象，所以我们只能选择第三种方法，即“ 精确 ”方法，默认时间限制在 5 分钟以内。一般认为样本量大于 25 可视为接近正态分布。



图 7.21 “ 精确检验 ”对话框

06 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TEST
   /WILCOXON=a 组--b 组  WITH  (PAIRED)
   /MISSING ANALYSIS
   /METHOD=EXACT TIMER(5).
```

07 单击“确定”按钮，执行操作。

执行完“两个相关样本检验”后，得到检验结果。

- 描述统计量

表 7.31　描述性统计量

Ranks

		N	Mean Rank	Sum of Ranks
脑中风-老年痴呆	Negative Ranks	5[a]	4.50	22.50
	Positive Ranks	2[b]	2.75	5.50
	Ties	0[c]		
	Total	7		

a．脑中风<老年痴呆

b．脑中风>老年痴呆

c．脑中风=老年痴呆

SPSS 首先输出两样本的等级信息，把两变量的差值的绝对值进行等级排列。表 7.31 中的“Nagative Ranks”行显示：A 组小于 B 组的差值共有 5 个，差值绝对值的等级平均值为 4.5，等级之和为 22.5。

- 统计检验结果

表 7.32　统计检验结果

Test Statistics[b]

	脑中风-老年痴呆
Z	-1.442[a]
Asymp.Sig.(2-tailed)	.149
Exact Sig.(2-tailed)	.188
Exact Sig.(1-tailed)	.094
Point Probability	.031

a．Based on positive ranks

b．Wilcoxon Signed Ranks Test

表 7.32 列出了统计检验结果。结果显示：使用“精确”方法计算的双侧显著性水平 ρ 值为 0.188，大于 0.05，所以两组患者的识字成绩差异不显著；也就是说，两个样本来自于同一总体，具有相同的总体分布。

7.8 K 个相关样本的非参数检验

SPSS 提供了三种检验多个相关样本的非参数检验方法：Friedman、Cochran's Q 和 Kendall's W。三种检验方法中，Friedman 又称弗里德曼双向等级方差分析，是 Wilcoxon 检验法的扩展。Cochran's Q 只适用于几个相关的二分变量（dichotomous variables），是 McNemar 方法的扩展。Kendall's W 方法又称肯德尔和谐系数，用于检验不同评价者的意见是否一致。

Friedman 和 Cochran's Q 都适用于重复测量设计和配对样本设计。在重复测量设计中，每个被试对象必须接受 k 个试验处理。相应地，每个被试对象在 k 个变量上都有得分，检验的目的在于发现被试对象在不同条件下是否有显著变化。在配对样本设计中，如果共有 k 个实验条件，则被试对象需要分为若干个小组，每个小组有 k 个被试对象，每一组的每个被试对象只接受一种实验处理，检验的目的在于发现不同实验条件下被试对象的得分是否有显著差异。

当使用 Kednall's W 时，SPSS 数据文件中的每一行代表一个评价者给出的评分，每一类代表一个被评价者获得的分数，W 系数最小为 0，表示几个评价者的意见不一致，最大为 1，表示几个评价者的意见完全一致。

如果检验结果发现样本间存在显著差异，则有必要做进一步的检验，例如 Wilcoxon 检验法，以确定具体差异所在。

多相关样本的非参数检验需要满足以下的前提假设条件。

- 随机性与独立性

每个（每组）被试对象的 k 个观测值必须代表随机样本，而且相互之间保持独立。

- 样本量

样本量越大，检验的精确度越高，当样本量超过 30 时，x^2 值会相当精确。

- 连续性和对称性（只适用于 Friedman）

任意两个变量差值的分布总体呈连续分布，而且有对称性。

【示例 7.8】多相关样本的非参数检验实例。

研究者进一步改进对脑损伤患者的研究方法。他们考察了 9 个额叶损伤病人对不同类别物体的图片和名称的匹配情况。物体分为三个类别：动物、水果和交通工具。给被试对象呈现物体图片和物体名称，要求被试对象将它们一一匹配起来。这 9 个病人在三类物体的得分如下表所示。

表 7.33　三类物体测试得分表

物体类别	得　分								
动物	6	4	7	11	5	5	7	12	8
水果	3	6	5	9	4	6	2	8	9
交通工具	7	5	4	6	8	3	6	8	10

试问这些病人在不同物体类别的得分有没有差异。

该问题的具体操作步骤如下。

01 建立数据文件。

将数据文件导入 SPSS 中，如图 7.22 示。

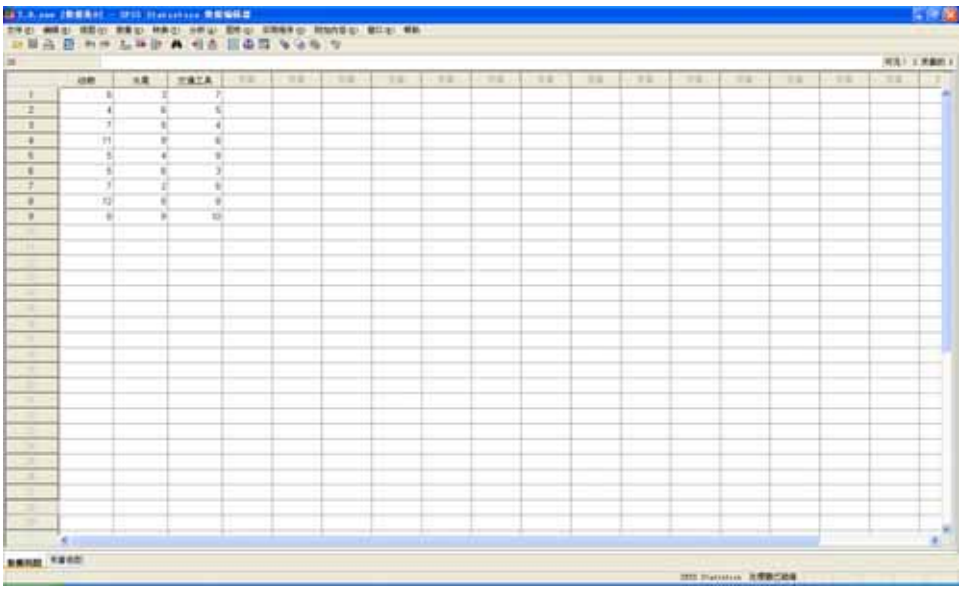


图 7.22　将数据文件导入 SPSS

通过分析发现，该数据包含三个变量：患者对三类物体的判别得分，分别命名为“ 动物 ”、“ 水果 ” 和 “ 交通工具 ”。其数据结构与单因素重复测量因素的结构相同。

02 打开检验对话框。

依次单击菜单“ 分析→非参数检验→K 个关联样本 ”命令，进入“ 多个关联样本检验 ” 对话框，如图 7.23 所示。



图 7.23 “ 多个关联样本检验 ” 对话框

03 确定检验变量。

依次选中图 7.23 左侧变量列表框中“ 水果 ”、“ 动物 ”和“ 交通工具 ”变量，单击 按钮，将其移动到 “ 检验变量 ” 列表框中，作为检验变量。

04 确定检验方法。

在 “ 多个相关样本检验 ” 对话框的 “ 检验类型 ” 选项组中选择 “ Friedman ” 检验法。

05 确定显著性水平计算方法。

在图 7.23 所示对话框中，单击 “ 精确 ” 按钮，打开 “ 精确检验 ” 对话框，如图 7.24 所示。本例我们只能选择第三种方法，即 “ 精确 ” 方法，默认时间限制在 5 分钟以内。

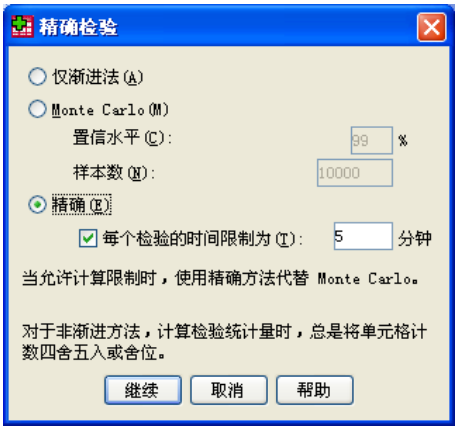


图 7.24 “ 精确检验 ” 对话框

06 单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
NPAR TESTS
   /FRIEDMAN = 动物 水果 交通工具
   /MISSING LISTWISE
   /METHOD=EXACT TIMER(5).
```

07 单击 “ 确定 ” 按钮，执行操作。

执行完 “ 多个相关样本检验 ” 后，得到检验结果。

- 描述统计量

表 7.34 描述性统计量

Ranks

	Mean Rank
动物	2．22
水果	1．83
交通工具	1．94

表 7.34 中显示了各组样本的均值，区分 “ 动物 ” 的平均得分为 2.22，区分 “ 水果 ” 的平均得分为 1.83，区分 “ 交通工具 ” 的平均得分为 1.94。

- 统计检验结果

表 7.35　统计检验结果

Test Statistics[a]

N	9
Chi-Square	.743
df	2
Asymp.Sig.	.690
Exact Sig.	.741
Point Probability	.087

表 7.35 所示的是 Fridman 检验统计结果，该结果表明患者对三类物体的识别成绩没有显著差异，“ 精确 ” 计算的显著性水平 ρ 值为 0.741，大于显著性检验水平 0.05，故拒绝接受零假设。

7.9　本章小结

本章讨论了非参数检验问题。非参数检验问题的方法较多，处理方式也不尽相同。从样本的角度可分为单样本的非参数检验和两独立样本的非参数检验问题；就具体方法而言，主要包括卡方检验、二项分布检验、游程检验、单样本 K-S 检验、两个独立样本检验、多独立样本检验、两个相关样本检验和多个相关样本检验等方法。读者应重点掌握非参数检验的基本思想，以及统计结果输出的实际含义。

第8章 相关分析

进行统计分析时，常常需要讨论两个或多个变量之间的相互关系。如果用户希望了解某个变量对另一变量的影响程度，则需要用到回归分析（见第9章）。如果用户希望了解两个或多个变量之间相互联系的密切程度，则需要相关分析来实现。通过相关分析，用户可以清晰地看出变量间的相互关系，从而判断出它们是否存在联系。

本章学习要点：

- 简单相关分析；
- 散点图；
- 偏相关分析。

8.1 简单相关分析

双变量相关分析，即简单相关分析，主要用于进行两个或多个变量间的相关分析（包括参数和非参数）。由于相关分析的绝大多数都通过两个变量进行相关分析，故有时直接将其简称为相关分析。如果进行两个变量的相关分析，则直接给出相关结果；如果进行多个变量的相关分析，则系统将给出多个变量间两两相关的分析结果。

在数理统计分析过程中，相关分析的使用非常广泛。例如，通过相关分析，可以判断学生的数学成绩和英语成绩高低是否相关。又例如，相关分析可以帮助实现验证家庭收入、消费、储蓄之间的关系。相关分析提供了衡量变量之间相互线形关系强弱的工具。

进行简单相关分析方法有两种：第一种，通过散点图直观显示变量之间关系（见第10章）；第二种，通过相关系数准确反映两变量的关系程度。两种方法各有优劣。利用散点图直观显示，相关关系更为简单明了、但不够精确；而利用相关系数反映正好相反，它能够以数字准确描述变量间的线性相关程度。

8.1.1 简单相关分析的基本思想

在介绍具体的简单相关分析操作步骤之前，我们首先要了解简单相关分析的理论基

础，即相关系数的统计含义。简单相关分析主要分两步进行：第一步，计算相关系数；第二步，利用假设检验进行统计推断。

在统计学中，一般用样本相关系数 r 推断总体相关系数 ρ。

$r=1$，两个变量总体之间完全正线性相关，一个变量变化，另一变量随之发生同比例的相同方向的变化。

$0.8\leq r<1$，两个变量总体之间呈高度正线性相关，两者的线性关系较强。

$0.3\leq r<0.8$，两个变量总体之间呈中度正线性相关，两者的线性关系中等。

$0<r<0.3$，两个变量总体之间呈轻度正线性相关，两者的线性关系较弱。

$r=0$，两个变量总体之间完全无关，即一个变量变化，另一变量不随之发生任何变化。

$-0.3\leq r<0$，两个变量总体之间呈轻度负线性相关，两者的线性关系较弱。

$-0.8\leq r<-0.3$，两个变量总体之间呈中度负线性相关，两者的线性关系中等。

$-1<r<-0.8$，两个变量总体之间呈高度负线性相关，两者的线性关系较强。

$r=-1$，两个变量总体之间呈完全负线性相关，一个变量变化，另一变量随之发生同比例的相反方向的变化。

通过样本相关系数 r，用户可以大体上推断出总体相关系数 ρ，并继而推断变量总体间的关系。然而，由于抽样过程受样本随机性、样本个数等多方面的影响，有时较大的样本相关系数并不能说明样本数据来自的两总体之间也存在显著的相关性；此时，就需要对其进行统计推断。简单相关分析的统计推断也采用了假设检验的方法。根据假设检验的基本思想，简单相关分析统计推断的基本步骤如下：

01 提出零假设 H_0：两总体之间不存在显著的线性相关，备择假设 H_1：两总体之间存在显著的线性相关；

02 根据不同的相关系数计算方法，计算相应的统计量；

03 根据统计量得到所对应的相伴概率；

04 用户给定一个显著性水平 α。如果相伴概率小于或等于显著性水平 α，则拒绝零假设 H_0，接受 H_1，认为两总体之间存在显著的线性相关。否则，如果相伴概率大于显著性水平 α，则接受零假设 H_0，认为两总体之间不存在显著的线性相关。

8.1.2　相关系数

根据不同的数据类型，需要采用不同的相关系数度量变量之间的线性相关关系。比较常用的相关系数有 Pearson 简单相关系数、Spearman 等级相关系数和 Kendall tua-b 一致性相关系数。下面分别就这三种相关系数进行介绍。

- Pearson 简单相关系数

Pearson 简单相关系数主要适用于度量定距型变量间（如 X、Y）的线性相关关系，不适用于度量非线性相关关系。其计算公式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{8.1}$$

其中，n 为样本容量。

Pearson 简单相关系数的对应检验统计量是 t 统计量，SPSS 将自动进行计算。t 统计量的计算公式为：

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\sim t(n-2) \tag{8.2}$$

t 统计量服从自由度为 n-2 的 t 分布，SPSS 将根据 t 统计量和自由度，依照 t 分布表自动给出 t 统计量所对应的相伴概率。如果相伴概率小于或等于显著性水平 α，则拒绝零假设 H_0；否则，接受零假设 H_0。

- Spearman 等级相关系数

Spearman 等级相关系数主要适用于度量定序变量间的线性相关关系。通过对两变量数据的秩（U_i、V_i）进行计算秩的方差，进而得到 Spearman 等级相关系数。其计算公式为：

$$r_s=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{8.3}$$

其中，n 为样本容量。

$$\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{n}(U_i-V_i)^2 \tag{8.4}$$

在对 Spearman 等级相关系数进行统计检验时，如果样本容量 n 小于或等于 30，则 SPSS 将依照 Spearman 等级相关统计量表给出相关系数对应的相伴概率。如果样本容量 n 大于 30，SPSS 将自动计算 Spearman 等级相关系数的对应检验统计量（Z 统计量）。Z 统计量的计算公式为：

$$Z=r_s\sqrt{(n-1)}\sim N(0,1) \tag{8.5}$$

Z 统计量近似服从标准正态分布，SPSS 将依照正态分布表自动给出 Z 统计量所对应的相伴概率。如果相伴概率小于或等于显著性水平 α，则拒绝零假设 H_0；否则，接受零假设 H_0。

- Kendall tua-b 一致性相关系数

Kendall tua-b 一致性相关系数同样适用于度量定序变量之间的线性相关关系。通过对两变量秩计算一致对（即同步增大）数目 U 和非一致对（即异序对）数目 V，进而得到 Kendall 的一致相关系数，其计算公式为：

$$\tau = (U - V)\frac{2}{n(n-1)} \tag{8.6}$$

其中，n 为样本容量，U 为一致对的数目，V 为非一致对的数目。

在对 Kendall tua-b 一致相关系数进行统计检验时，如果样本容量 n 小于或等于 30，则 SPSS 将依照 Kendall tua-b 一致相关统计量表给出相关系数对应的相伴概率。如果样本容量 n 大于 30，SPSS 将自动计算 Kendall tua-b 一致相关系数的对应检验统计量（Z 统计量）。Z 统计量的计算公式为：

$$Z = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \sim N(0,1) \tag{8.7}$$

Z 统计量近似服从标准正态分布，SPSS 将依照正态分布表自动给出 Z 统计量所对应的相伴概率。如果相伴概率小于或等于显著性水平 α，则拒绝零假设 H_0；否则，接受零假设 H_0。

8.1.3　简单相关分析的操作步骤

01 打开数据文件，选择“分析（Analyze）”菜单下的“相关（Correlate）”子菜单，如图 8.1 所示。



图 8.1　“分析”菜单下的“相关”子菜单

单击“双变量（Bivariate）”命令，SPSS 将弹出“双变量相关（Bivariate Correlations）”

主对话框，如图 8.2 所示。



图 8.2 “双变量相关（Bivariate Correlations）”主对话框

在“双变量相关（Bivariate Correlations）”主对话框中，用户可以从左边源变量列表中选择两个或两个以上需要进行相关分析的变量，单击 进入右框的“变量（Variables）”列表框中。

相关系数（Correlation Coefficients）栏：用户可以选择 Pearson（简单相关系数）、Spearman（等级相关系数）或 Kendall 的 tua-b（一致性相关系数）作为计算的相关系数。

显著性检验（Test of Significance）栏：用户可以选择相关系数的假设检验的检验方式。

双侧（Two-tailed）检验：如果用户检验前不确定两变量之间是正相关还是负相关，则选择双侧检验。SPSS 默认为双侧检验。

单侧（One-tailed）检验：如果用户检验前已经确定两变量之间是正相关或负相关，则选择单侧检验。

标记显著性相关（Flag significant correlations）选项：用户可以选择是否在相关分析的结果中显示统计检验的相伴概率，相伴概率值将以星号（*）标记。如果用户指定的显著性水平为 0.05，当统计量的相伴概率小于或等于 0.05 时，相关系数值旁将标识一个星号（*）。如果用户指定的显著性水平为 0.01，当统计量的相伴概率小于或等于 0.01 时，相关系数值旁将标识两个星号（**）。显著性水平为 0.01 时，即两个星号（**）时，统计检验更为精确。

02 单击“选项（Options）”按钮，打开“双变量相关性：选项（Bivariate Correlations：Options）”对话框，如图 8.3 所示。

在“双变量相关性：选项（Bivariate Correlations：Options）”对话框中，用户可以指

定输出分析结果和缺失值的处理方式。

统计量（Statistics）栏：选择是否输出及输出的基本描述统计量。该栏只有在选择当且仅当计算 Pearson（简单相关系数）时，才能进行选择。

均值和标准差：将输出每个变量的均值和标准差。

叉积偏差和协方差：将输出每对变量的交叉积离差和协方差。交叉积离差即为 Pearson 简单相关系数计算公式的分子部分，协方差等于交叉积离差/（n-1）。两者均为反映变量间相关程度的描述统计量。

缺失值（Missing Values）栏：用户可以定义分析中对缺失值的处理方式，包括“ 按对排除个案 ”和“ 按列表排除个案 ”。为尽可能充分利用数据，SPSS 默认选择前者。

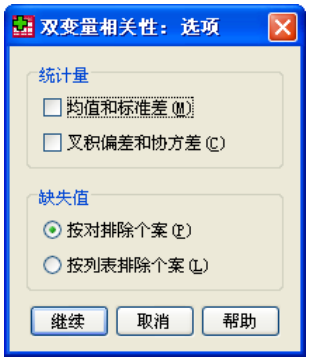


图 8.3“ 双变量相关性 :选项（Bivariate Correlations：Options）”对话框

指定输出分析结果和缺失值的处理方式后，单击“ 继续 ”按钮，即可返回到“ 双变量相关 （Bivariate Correlations）”主对话框中。

03 在“ 双变量相关（Bivariate Correlations）”主对话框中，单击“ 确定（OK）”按钮，即可在结果输出窗口得到相关分析的结果输出表格。相关分析的结果输出表格主要包括以下两种：

1）基本统计描述（Descriptive Statistics）列表：如果用户在“ 双变量相关性：选项（Bivariate Correlations：Options）”对话框中选择了统计量（Statistics），则输出该报表。SPSS 将根据用户选择的不同，输出平均值和标准差或者叉积偏差和协方差。

2）相关系数检验（Correlations）：以方阵的形式列出两两变量间的相关系数。表格中，每一行和每一列的两个变量所对应的格子即为这两变量相关分析结果。相关分析结果共分为三项，包括相关系数（Correlation）、相伴概率（即 P 值，Sig.）和样本个数（N）。同时，如果用户在“ 双变量相关（Bivariate Correlations）”主对话框中选中标记显著性相关（Flag significant correlations）选项，那么在满足条件的相关系数旁将标识星号。

8.2　散点图

除了利用相关系数来表示两个变量之间的线性相关关系外，SPSS 还提供了散点图来更为直观地显示变量之间的相关关系。通过绘制散点图，用户可以判断变量之间是否存在相关关系，并且可以根据散点的分布，大体估计出变量之间符合的曲线类型。绘制散点图的具体操作步骤如下。

01 打开数据文件，选择“ 图表（Graphs）”菜单，单击“ 旧对话框 ”下的“ 散点图（Scatter/Dot）”。SPSS 将弹出“ 散点图/点图（Scatter/Dot）”对话框，如图 8.4 所示。

SPSS 提供了下面 5 种类型的散点图，满足不同用户的需求。

简单分布（Simple Scatter）：显示一对变量之间的散点图。以一个变量的观测值为横轴、以另一个变量的观测值为纵轴来确定散点的坐标。



图 8.4 “ 散点图/点图（Scatter/Dot）” 对话框

矩阵分布（Matrix Scatter）：以矩阵的形式，在多个坐标轴上显示多对变量的散点图。矩阵的行列上的格子数是相等的，正对角线的格子中显示参加绘图的若干个变量的名称。每个变量名所在的行和列决定了矩阵分布图各个图形的行列坐标的含义。

简单点（Simple Dot）：以一维形式对变量的散点进行纵向叠加。

叠加分布（Overlay Scatter）：在同一坐标轴上，显示多对变量的散点图，即将多张简单分布图集中在一个坐标轴上显示。

3-D 分布（3-D Scatter）：以立体图的形式，在立体空间中显示 *X*、*Y*、*Z* 三个变量间的散点图。

下面以两个变量间的相关分析为例，绘制散点图。

02 选中 “ 简单分布（Simple Scatter）” 命令，单击 “ 定义（Define）” 按钮，打开 “ 简单散点图（Simple Scatterplot）” 对话框，如图 8.5 所示。

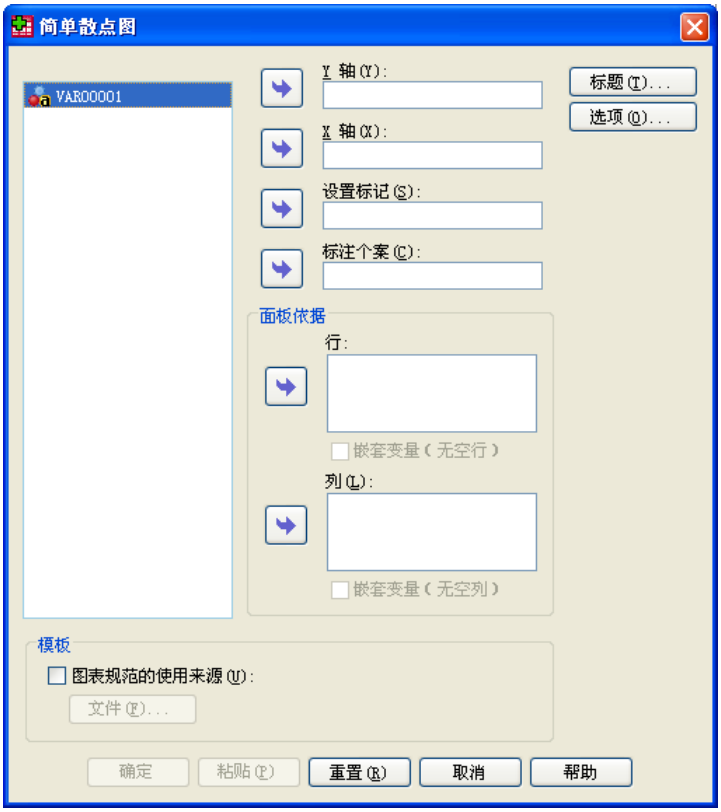


图 8.5 “ 简单散点图（Simple Scatterplot）” 主对话框

在“简单散点图（Simple Scatterplot）”对话框中，用户可以从左边源变量列表中选择变量，单击进入右框，定义图表。

- ***Y* 轴（Y Axis）**：用于定义散点图的 *y* 轴。
- ***X* 轴（X Axis）**：用于定义散点图的 *x* 轴。
- **设置标记（Set Markers by）**：该列表框用于指定分组变量。如果指定了分组变量，SPSS 会按照分组变量的不同取值，将观察值分成几组，然后对不同的分组分别以不同的颜色绘制散点图。
- **标注个案（Label Cases by）**：该列表框用于指定标记变量。如果指定了标记变量，SPSS 会在散点图中的每个点上标出该个案的相应标记变量的数值。
- **模板（Template）**：用户可以选中该选项，并单击“文件（File）”按钮，选择模板文件，作为散点图的格式模板。

03 单击“标题（Titles）”按钮，打开“标题（Titles）”对话框，如图 8.6 所示。

在“标题（Titles）”对话框中，用户可以定义散点图的标题、子标题和脚注。

- **标题（Title）和子标题（Subtitle）**：通过输入，用户可以定义两行标题和一行子标题。标题和子标题将在散点图上方显示。用户也可以在结果输出窗口中定义标题。
- **脚注（Footnote）**：通过输入，用户可以定义两行脚注，脚注将在散点图下方显示。

输入标题后，单击“继续”按钮，即可返回“简单散点图（Simple Scatterplot）”主对话框中。

04 单击“选项（Options）”按钮，打开“选项（Options）”对话框，如图 8.7 所示。

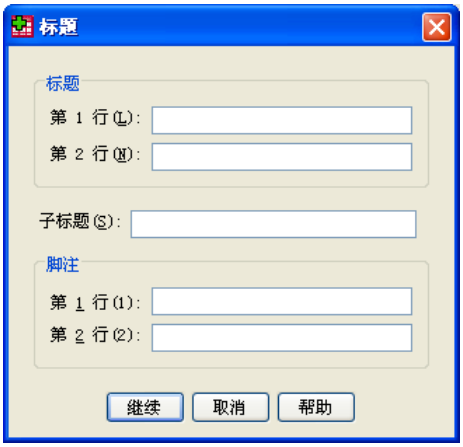


图 8.6 “标题（Titles）”对话框

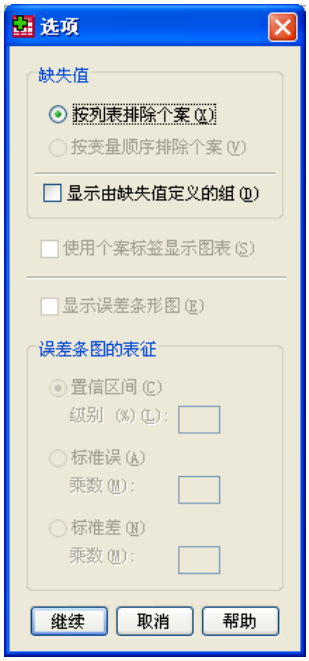


图 8.7 “选项（Options）”对话框

在“选项（Options）”对话框中，用户可以指定缺失值的处理方式和误差条的设定形式。

缺失值（Missing Values）栏：用户可以定义分析中对缺失值的处理方式，包括“按变量顺序排除个案”和“按列表排除个案”。

误差条图的表征（Error Bars Represent）栏：用户可以选择设定置信区间、标准误差或者标准偏差条件，并显示误差条图。

指定缺失值的处理方式和误差条的设定形式后，单击“继续”按钮，即可返回“简单散点图”主对话框中。

05 在“简单散点图（Simple Scatterplot）”主对话框中，单击“确定（OK）”按钮，即可在结果输出窗口得到简单散点图。

8.3 偏相关分析

简单相关关系只能反映两个变量之间的线性相关关系，但如果当因变量受到多个因素影响时，因变量与其中某一个自变量之间的简单相关关系显然也会受到其他相关因素的影响。此时，简单相关关系就不能真实地反映两者之间的关系。

剔除其他因素的影响后，只考察两者之间的相关程度，所做的分析即为偏相关分析。在偏相关分析中，能够人为控制的影响因素（或变量）称为控制变量。与简单相关分析相似，在进行偏相关分析时，也将计算一个系数表示两个变量之间的相关程度，这个系数称为偏相关系数。

假设存在自变量 x_1、x_2 和因变量 y，其中，x_2 作为控制变量。那么 x_1 和 y 之间的偏相关系数为：

$$r_{y1,2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} \tag{8.8}$$

其中，r_{y1} 表示 x_1 和 y 之间的简单相关系数，r_{y2} 表示 x_2 和 y 之间的简单相关系数，r_{12} 表示 x_1 和 x_2 之间的简单相关系数。

偏相关分析的具体操作步骤如下。

01 打开数据文件，选择“分析（Analyze）”菜单，单击“相关（Correlate）”下的“偏相关分析（Partial）”命令。SPSS 将弹出“偏相关（Partial Correlations）”对话框，如图 8.8 所示。

在“偏相关（Partial Correlations）”对话框中，用户可以从左边源变量列表中选择两个或两个以上需要进行偏相关分析的变量，单击 ▸ 进入右框的“变量（Variables）”列表框中。同时，用户需要选择一个或多个控制变量，单击 ▸ 进入右框的“控制（Controlling）”列表框中。如果不选择控制变量，则 SPSS 进行的就是简单相关分析。

显著性检验（Test of Significance）栏：用户可以选择偏相关系数的假设检验的检验方式。



图 8.8 “ 偏相关（Partial Correlations）” 主对话框

双侧（Two-tailed）检验：如果用户检验前不确定两变量之间是正相关还是负相关，则选择双侧检验。SPSS 默认为双侧检验。

单侧（One-tailed）检验：如果用户检验前已经确定两变量之间是正相关或负相关，则选择单侧检验。

显示实际显著性水平（Display actual significance level）选项：用户可以选择是否在偏相关分析的结果中显示给出确定的相伴概率（即 P 值），SPSS 系统默认为选中。

02 单击“ 选项（Options）”按钮，打开“ 偏相关性 :选项（Partial Correlations :Options）” 对话框，如图 8.9 所示。

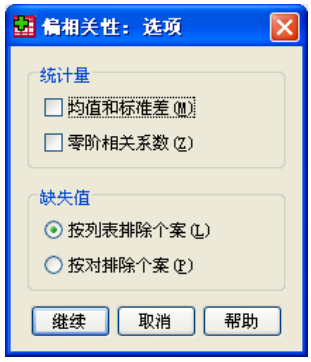


图 8.9 “ 偏相关性：选项（Partial Correlations：Options）” 对话框

“ 偏相关性 :选项（Partial Correlations :Options）”对话框与“ 双变量相关 :选项（Bivariate Correlations：Options）” 对话框大体相同，只是在输出的描述统计量上，将叉积偏差和协方差（Cross-product deviations and covariances）转换成了“ 零阶相关系数（Zero-order

Correlations）”。如果选择该项，则在计算输出变量间的偏相关系数的同时，还输出变量间的简单相关系数。其他相同的选项；这里不再赘述。

03 在“偏相关（Partial Correlations）”主对话框中，单击“确定（OK）”按钮，即可在结果输出窗口得到偏相关分析的结果输出表格。偏相关分析的结果输出表格主要包括以下两种：

1）基本统计描述（Descriptive Statistics）列表：如果用户在“偏相关：选项（Partial Correlations：Options）”对话框中选择了描述性统计（Statistics），则输出该报表。SPSS 将根据用户选择的不同，输出平均值和标准差或者零次序相关。

2）偏相关分析输出表（Partial Correlation Coefficients）：与相关分析的输出表格相同，以方阵的形式列出两两变量间的偏相关系数。表格中，每一行和每一列的两个变量所对应的格子即为这两个变量偏相关分析结果。相关分析结果共分为三项，包括偏相关系数（Correlation）、自由度（df）和相伴概率（即 P 值，Sig.）。

8.4 本章小结

本章主要介绍了相关分析的基本思想，包括简单相关分析和偏相关分析，对应于“分析（Analyze）”菜单下的“相关（Correlate）”子菜单（如图 8.1 所示）。通过本章的学习，读者可以判断数据中任意两个变量间是否存在线性相关，并且能够剔除控制变量，进行偏相关分析。相关分析在数理统计分析中较为常用，内容也比较简单，在相关分析的基础上，我们将在第 9 章深入研究线性回归分析。

第9章 回归分析

回归分析（Regression）是一种应用极为广泛的统计分析方法，已成功应用在金融经济学、医学等领域。它用于分析事物之间的统计关系，侧重考察变量之间的数量变化规律，并通过回归方程的形式描述和反映这种关系。

回归分析的方法有多种，本章重点讨论回归分析的基本思想和多种回归分析方法的 SPSS 实现过程。在 SPSS 中，回归分析模块提供了回归分析的多种重要方法，主要包括以下过程：

- 线性回归；
- 曲线估计；
- 二元 logistic 回归；
- 多项 logistic 回归；
- 有序；
- Probit 分析；
- 非线性；
- 权重估计；
- 两阶最小二乘法；
- 最佳尺度。

在 SPSS 中可以利用“分析→回归分析”命令实现回归分析，菜单内容如图 9.1 所示。

本章学习要点：

- 线性回归；
- 非线性回归；
- 曲线估计；
- logistic 回归。

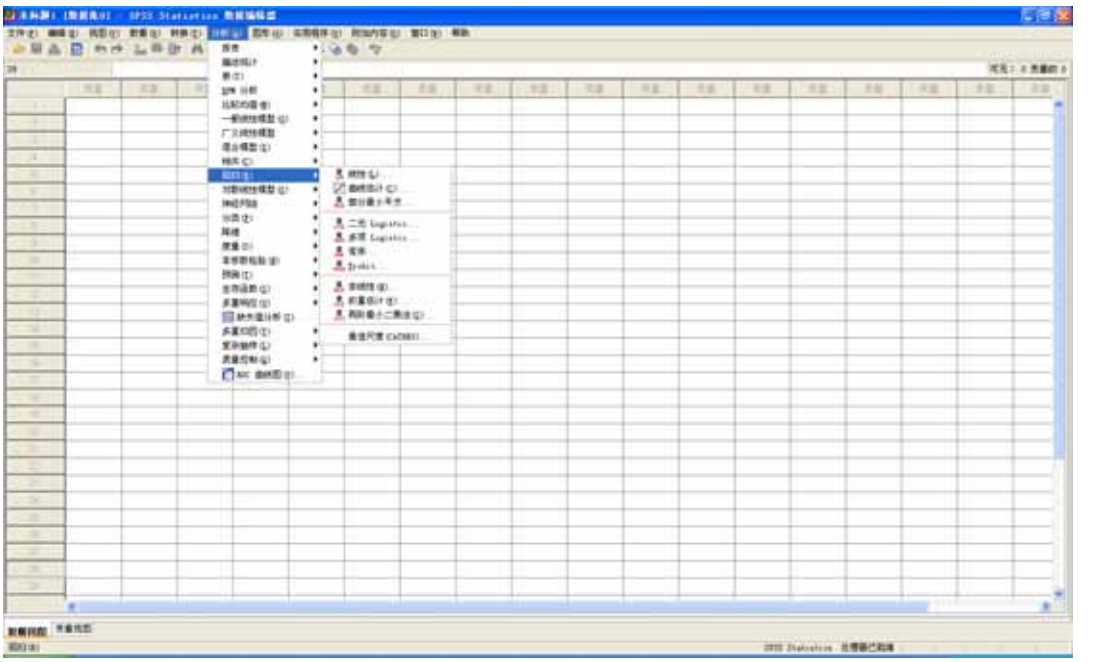


图 9.1　SPSS 回归分析模块

9.1　线性回归

线性回归分析（Linear Regression）是研究一个因变量和一个或多个自变量之间是否存在某种线性关系的统计学方法。如果参与回归分析的自变量只有一个，就是线性回归分析，也称为直线回归分析，得到的结果称为直线回归方程。如果参与回归分析的变量有多个，则是多元线性回归。

9.1.1　线性回归的基本原理

本节以一元线性回归分析为例，讨论线性回归的基本思想和回归模型的分析过程。对于多元线性回归分析，基本思想与一元线性回归分析类似。

线性回归分析是基于最小二乘原理的统计分析方法，是在统计假设下的最优线性无偏估计。

下面以一元线性回归分析为例，说明数学模型的建立过程。

假设已经收集到 n 对数据 $(x_i, y_i)(i=1,2,...n)$，这里 x_i 可以认为是来自于随机变量 x 的一组样本值，y_i 被认为是来自于 y 的另一组样本值；当讨论 x 和 y 之间是否存在统计意义下的相互关系时，x 和 y 又称为解释变量和被解释变量。

一元线性回归的数学模型为：

$$y = a + bx + \varepsilon \tag{9.1}$$

在（9.1）式中，被解释变量 y 的变化可由下列两个部分解释：

- 解释变量 x 的变化引起的 y 的线性变化部分，即 $y = a + bx$ ；
- 其他随机因素引起的 y 的变化部分，即 ε。

a 和 b 都是模型的未知参数，分别称为回归常数和回归系数，ε 是回归模型的随机误差，满足如下条件：

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = \delta^2 \end{cases} \tag{9.2}$$

9.2 式表明，随机误差的数学期望为 0，随机误差的方差是一个特定的常数。如果对回归模型求期望，则得：

$$E(y) = a + bx \tag{9.3}$$

进行先行回归分析的前提是根据给定的数据对估计线性方程的位置参数 a 和 b，利用最小二乘法可以得到参数的无偏估计，由于参数的估计值是根据样本值得到的，因此由此得到的数据值只是真值 a 和 b 的估计值，记为 $\hat{a}$ 和 $\hat{b}$。根据参数的估计值可以定义一元线性经验回归方程：

$$\hat{y} = \hat{a} + \hat{b}x \tag{9.4}$$

一元线性经验回归方程在二维平面上表示为一条直线，即回归直线。它表示当解释变量 x 变化时引起的被解释变量 y 的变化的估计值。

利用最小二乘法总能够计算出线性回归模型中的参数，但由此确定的线性回归方程不能立即用于对实际问题的分析，还必须对回归方程的线性关系进行各种统计检验，包括回归方程的显著性检验、回归系数的显著性检验、残差分析等。

线性回归方程显著性检验的零假设为 $H_0 : b = 0$，即检验回归参数是否为零；如果为零，说明被解释变量和解释变量之间不具有线性关系，回归方程没有意义，线性回归方程不能够解释被解释变量和解释变量之间的关系。

下面给出几种方法定义的统计量。

- F 检验法

该检验方法定义的统计量为：

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2)} \tag{9.5}$$

在式 9.5 中，$\bar{y}$ 表示被解释变量 y 的样本均值，统计量 F 服从自由度为 $(1, n-2)$ 的 F 分布。在 SPSS 操作中，将根据检验统计量自动计算检验统计量的观测值和对应的概率 ρ 值。

- T 检验法

T 检验法的检验统计量定义为：

$$T=(\sum_{i=1}^{n}(x_i-\overline{x})^2)^{\frac{1}{2}}\frac{\hat{b}}{\hat{\delta}} \tag{9.6}$$

其中，(9.6) 式中的 $\hat{\delta}$ 为 δ 的无偏估计量，计算公式为 $\hat{\delta}=\sqrt{\frac{1}{n-2}(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2)}$。统计量 T 服从自由度为 $n-2$ 的 t 分布。在 SPSS 操作中，将根据检验统计量自动计算统计量的观测值和对应的概率 ρ 值。

- 相关系数检验法

相关系数检验法的检验统计量定义为：

$$R=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2)^{\frac{1}{2}}} \tag{9.7}$$

(9.7) 式中的统计量 R 称为相关系数，在 SPSS 操作中，将根据检验统计量自动计算统计量的观测值和对应的概率 ρ 值。

残差是指由回归方程计算所得的预测值与实际样本值之间的差距，即：

$$e_i=y_i-\hat{y}_i=y_i-(\hat{a}+\hat{b}x_i) \tag{9.8}$$

如果回归方程能够较好地反映被解释变量的特征和变化规律，则残差中应该不包含明显的规律性和趋势性。基于这一点，残差分析的内容包括分析残差是否服从均值为 0 的正态分布、分析残差是否为等方差的正态分布、分析残差序列是否独立等。数值方法和图形直观分析方法都是分析过程中有效的分析工具。

检验残差序列是否独立，可以通过构造检验统计量进行检验。这里给出 Durbin-Watson 检验法，该检验法定义的检验统计量为：

$$DW=\frac{\sum_{e=2}^{e}(e_i-e_{i-1})^2}{\sum_{i=1}^{n}e_i^2} \tag{9.9}$$

(9.9) 式中 DW 统计量的取值范围在 0~4 之间。当 $|DW-2|$ 过大时拒绝零假设，认为两残差之间是相关的；当 $DW>2$ 时，认为两残差之间为正相关，当 $DW<2$，则认为两残差之间为负相关，只有 DW 近似为 2 时，才认为两残差之间是相互独立的。

残差的方差齐性检验可以通过生成回归标准化测量的散点图来检验。当图中各点分布没有明显的规律性，即残差的分布不随预测值的变化而增大或减小时，则认为残差满足方差齐性条件。

除此之外，还可以通过保存残差，对保存变量用方差齐性检验方法进行检验。

9.1.2　SPSS 线性回归

SPSS 中一元线性回归分析和多元线性回归分析的功能菜单是集成在一起的。下面给出 SPSS 操作步骤和对话框介绍。

- 选择线性回归对话框

依次单击菜单“ 分析→回归分析→线性 ”命令，打开“ 线性回归 ”对话框，如图 9.2 所示。

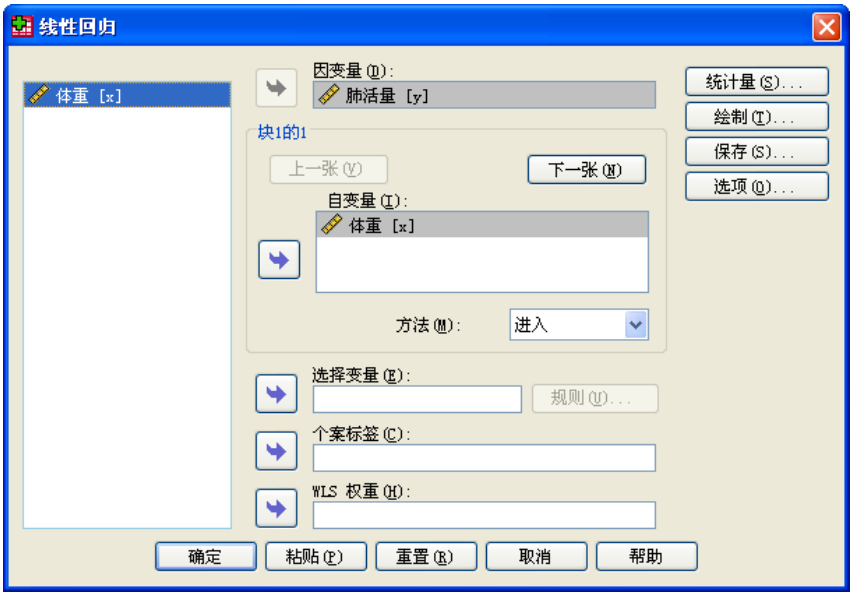


图 9.2 “ 线性回归 ”对话框

- 选择被解释变量和解释变量

在变量列表框中选择数值型变量后，单击 按钮，将其移动到“ 因变量 ”列表框中，作为线性回归分析的被解释变量；然后选择一个或多个变量，单击 按钮，将其移动到“ 自变量 ”列表框，作为线性回归分析的解释变量。

如果要对不同的因变量采用不同的引入方法，例如对某两个变量用强迫引入法，对其他自变量用前引入法，这时可利用“ 上一张 ”与“ 下一张 ”按钮把自变量归类到不同的自变量块中，然后对不同的变量子集选用不同的引入方法。

- 选择个案标签

在变量列表框中选择变量，单击 按钮，将其移动到“ 个案标签 ”列表框中，被选变量的标签用于在图形中标注点的值。

- 选择 WLS 权重标签

在变量列表框中选择变量，单击 按钮，将其移动到“ WLS 权重 ”列表框中，该选项仅在被选变量为权变量时选择。

● 执行操作

选择完毕后，单击“ 确定 ”按钮，执行线性回归分析的操作。

下面对“ 线性回归 ”对话框中的子对话框和各参数项进行简单的介绍。

● “ 方法 ”选项

单击“ 方法 ”下拉列表框中的三角箭头，选择建立线性回归方程的方法，包括 5 个选项，其内容及具体含义如表 9.1 所示。

表 9.1　方法选项及其含义

方　法	含　义
进入	默认选项。全部被选变量一次性进入回归模型，这是最简单的一种回归估计方法
逐步	在每一次引入变量时，概率 F 最小值的变量将引入回归方程。如果已引入回归方程的变量的 F 大于设定值，将被剔除出回归方程。当无变量被引入或剔除时，终止回归过程
删除	将所有不进入方程模型的被选变量一次性剔除
向后	一次性将所有变量引入方程，并依次进行剔除。首先剔除与因变量最小相关且符合剔除标准的变量，然后进行第二个与因变量最小相关并且符合剔除标准的变量，依次类推。当方程的变量均不满足剔除标准时，终止回归方程
向前	被选变量依次进入回归模型，首先引入与因变量最大相关且符合引入标准的变量，在引入第一个变量后，再引入第二个与因变量最大偏相关并且符合引入标准的变量，依次类推。当无变量符合引入标准时，终止回归过程

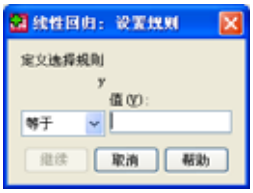


图 9.3 “ 线性回归：设置规则 ”对话框

● “ 规则 ”选项

在“ 线性回归 ”对话框的变量列表窗口中选择列表，将其移动到“ 选择变量 ”列表框，该变量用于指定分析个案的选择规则。单击“ 规则 ”按钮，打开“ 线性回归：设置规则 ”对话框，如图 9.3 所示。

在该对话框的“ 定义选择规则 ”中左侧的下拉菜单中给出设定临界值规则的选项。其内容及具体含义如表 9.2 所示。

表 9.2　方法选项及其含义

选　项	含　义
等于	等于选定值
不等于	不等于选定值
小于	小于选定值
小于或等于	小于或等于选定值
大于	大于选定值
大于或等于	大于或等于选定值

在“ 值 ”文本框中输入相因变量的设定规则的临界值。选择完毕后，单击“ 继续 ”按钮，返回“ 线性回归 ”对话框。

● “ 统计量 ” 选项

在如图 9.2 所示的 “ 线性回归 ” 对话框中单击 “ 统计量 ” 按钮，打开 “ 线性回归：统计量 ” 对话框，如图 9.4 所示。

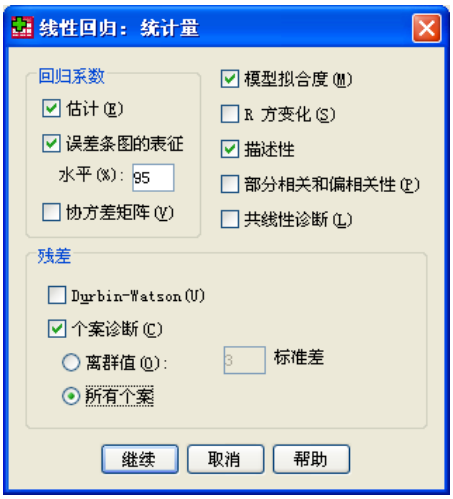


图 9.4 “ 线性回归：统计量 ” 对话框

在该对话框的 “ 回归系数 ” 选项组控制有关回归系数统计量的显示，其内容及其含义如表 9.3 所示。

表 9.3　回归系数选项及其含义

选　项	含　义
估计	默认选项。显示回归系数的估计值 β、回归系数的标准误差、标准化回归系数、回归系数的 β 的 t 估计值和双尾显著性水平
模型拟合度	列出进入或从模型中剔除的变量以及拟合优度统计量、复相关系数（R）、判定系数（R^2）、调整 R^2、估计值的标准误差以及方差分析表
误差条图的表征	显示回归系数的 β 的 95%置信区间
R 方变化	增加或删除一个因变量所产生的 R^2 改变量。R^2 改变量越大，表明该变量可能是一个较理想的回归变量
协方差矩阵	显示回归系数的方差—协方差系数矩阵
描述性	显示变量的均值、标准差、相关系数矩阵及单尾检验
部分相关和偏相关分析	显示零阶相关、偏相关和部分相关系数。方程中至少有两个自变量
共线性诊断	显示包括各变量的容差、方差膨胀因子和共线性的诊断表

“ 残差 ” 选项组中的选项用于进行有关残差的设置，其选项及其含义如表 9.4 所示。

表 9.4　残差选项及其含义

选　项	含　义
Durbin-Waston	显示残差相关的 Durbin-Waston 检验和残差与预测值的综述统计

续表

选　项	含　义
个案诊断	选择该项后，在下面的两个单选项中进行选择，确定几逆行能够残差分析的个案范围，选项包括“离群值”，含义为超过 n 倍标准差以上的个案为奇异值，默认值为 3；“所有个案”表示显示所有变量的标准化残差、观测值和预测值、残差

- “绘图”选项

在“线性回归”对话框中单击“绘制”按钮，打开“线性回归：图”对话框，如图 9.5 所示。



图 9.5　“线性回归：图”对话框

在“线性回归：图”对话框中可以分析数据资料的正态性、线性和方差齐性，还可以检测奇异值或异常值等，是研究回归分析中线性规律很有用的工具。

对话框左侧的列表框中包含选择绘制散点图的坐标轴变量，用户可以从列表框中选择变量作为 X（横轴变量）和 Y（纵轴变量）；还可以通过单击“下一张”按钮来重复操作过程，以获得更多的图形。列表项目及其含义如表 9.5 所示。

表 9.5　列表项及其含义

选　项	含　义
DEPENDENT	因变量
*ZPRED	标准化预测值
*ZRESID	标准化残差
*DRESID	删除残差
*ADJPRED	调整预测值
*SRESID	Student 残差
*SDRESID	Student 删除残差

在“标准化残差图”选项区域中选择将要生成的标准化残差图的类型，各选项及含义如表 9.6 所示。

表 9.6　标准化残差图选项及其含义

选　　项	含　　义
直方图	标准化残差的直方图，并给出正态曲线
正态概率图	可对标准化残差分布与正态分布进行比较
产生所有部分图	生成每个自变量残差与因变量残差的散点图，最少有两个自变量引入方程才能产生偏残差图

- “保存”选项

在“线性回归”对话框中单击“保存”按钮，打开“线性回归：保存”对话框，如图 9.6 所示。

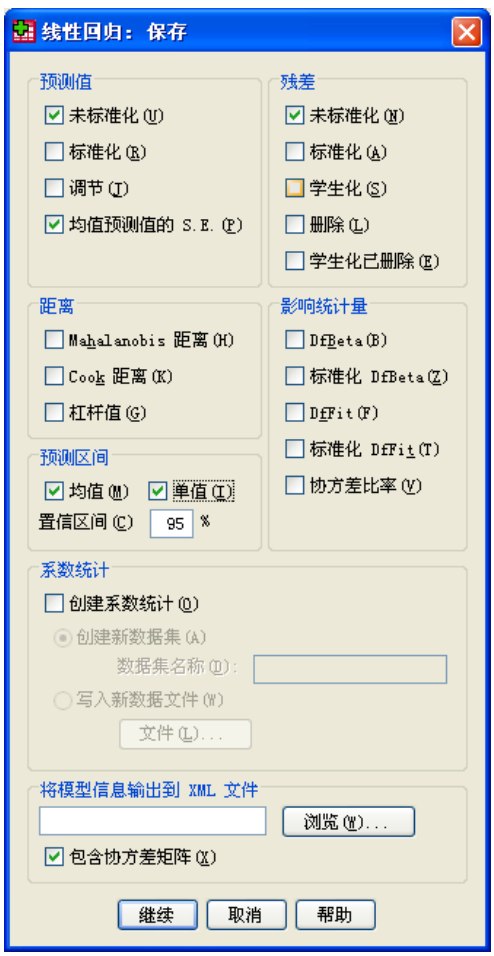


图 9.6　“线性回归：保存”对话框

利用该对话框，可以选择保存运行过程中生成的部分或全部统计量。对话框中各选项组的含义如表 9.7、表 9.8 和表 9.9 所示。

表 9.7　预测值选项组中选项及其含义

选　项	含　义
未标准化	非标准化的预测值。模型中因变量的预测值
标准化	标准化的预测值。将每个预测值转换成标准化形式
调节	调整预测值。在回归系数的计算中剔除当前个案的预测值
均值预测值的 S.E.	预测值均值的标准误差

表 9.8　距离选项组中选项及其含义

选　项	含　义
Mahalanobis 距离	是自变量个案值与所有个案平均值的距离，Maholanobis 距离大时，表示该个案的一个或多个自变量有奇异值
Cook 距离	当一个个案参与计算回归线系数时，所有个案残差变化的大小；Cook 距离大时，表示该个案直接影响回归系数
杠杆值	测量回归拟合点的影响，中心杠杆值的范围是 0～（*n-1*）/*n*

表 9.9　预测区间选项组中选项及其含义

选　项	含　义
均值	平均预测区间，表示平均预测区间与个体预测区间的上下限
单值	个体预测区间，对于单个自变量，因变量平均预测区间的上下限

在“预测区间”选项组选择了“均值”和“单值”后，“置信区间”文本框被激活，在该文本框中输入 1~99.9 之间的任意数值，用于指定上述两个预测区间的置信度。

“残差”用于设置残差的保存选项，有“未标准化”、“标准化”、“学生化”、“删除”和“学生化已删除”选项。

“影响统计量”中的选项反映了剔除某个案的数据以后回归系数的变化情况，选项及其含义如表 9.10 所示。

表 9.10　影响统计量选项组中选项及其含义

选　项	含　义
DfBeta	DfBeta 值，为删除一个个案后回归系数改变的大小。该值在模型中的每一项均进行计算
标准化 DfBeta	标准化 DfBeta 值，为删除一个个案后回归系数改变的大小。当它大于 2/SQRT(n)时，该点可能为强影响点。该值在模型中的每一项均进行计算
DfFit	DfFit 值，即拟合值之差，删除一个个案后回归系数改变的大小
标准化 DfFit	标准化 DfFit 值，拟合值的标准化差值，删除一个个案后回归系数改变的大小。当它大于 2/SQRT(n)时，该点可能为强影响点
协方差比率	协方差矩阵的比率，为一个个案不参加回归系数计算时的协方差矩阵与所有个案均参与回归系数计算协方差矩阵的比值。该值若接近于 1，表明该个案对协方差矩阵的影响不大

- “选项”对话框

在“线性回归”对话框中单击“选项”按钮，打开“线性回归：选项”对话框，如图 9.7 所示。

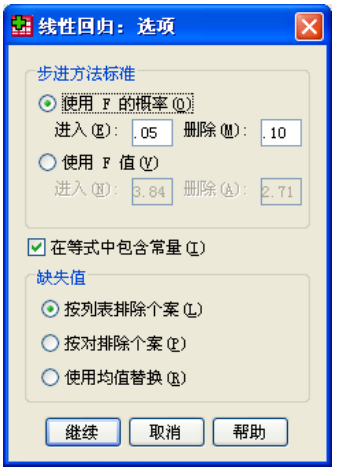


图 9.7 “ 线性回归：选项 ” 对话框

在 “ 线性回归：选项 ” 对话框中，“ 步进方法标准 ” 选项组给出利用逐步回归方法时进入值和删除值的标准，其选项及其含义如表 9.11 所示。

表 9.11　逐步方法标准选项组中选项及其含义

选　项	含　义
使用 F 的概率	使用 F 显著水平值。当候选变量中最大 F 值的概率 ρ 值小于或等于引入值时，引入相应的变量；在已进入方程的变量中，最小 F 值的概率 ρ 值大于或等于剔除值（默认为 0.10）时，剔除相应的变量。所设定的引入值必须小于剔除值，用户可以设定其他标准，如引入 0.10 剔除 0.11，放宽变量进入方程的标准
使用 F 值	使用 F 值，含义同上

在默认状态下，回归模型包含 “ 在等式中包含常量 ” 选项，若不选此项，则回归模型经过原点。回归方程是否包含常数项，对某些回归结果的解释将是不一样的。

“ 缺失值 ” 选项组中给出处理缺失值的方式，主要有 “ 按列表排除个案 ”、“ 按对排除个案 ” 和 “ 使用均值替换 ” 三种。

9.1.3　一元线性回归

【示例 9.1】体检数据中的体重和肺活量的分析

某单位对 12 名女工进行体检，体检项目包括体重（kg）和肺活量（L），数据如表 9.12 所示。

表 9.12　12 名女工体检数据

项　目	数　据
x（体重：kg）	42.00 42.00 46.00 46.00 46.00 50.00 50.00 50.00 52.00 52.00 58.00 58.00
y（肺活量：L）	2.55 2.20 2.75 2.40 2.80 2.81 3.41 3.10 3.46 2.85 3.50 3.00

利用 x 表示体重，y 表示肺活量，建立数据文件。利用一元线性回归分析描述其关系。具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开线性回归对话框和选入变量。

依次单击菜单“分析→回归分析→线性”命令，打开“线性回归”对话框。在“因变量”列表框中选择“肺活量”，作为线性回归分析的被解释变量；在“自变量”列表框中选择“体重”，作为解释变量。

03 设置。

单击“统计量”按钮，打开“线性回归：统计量”对话框，依次选择“估计”、“误差条图的表征”、“模型拟合度”、“描述性”、“个案诊断”和“所有个案”等选项。

单击“绘制”按钮，打开“线性回归：图”对话框，选择“DEPENDNT”作为 y 轴变量，“*ZPRED”作为 x 轴变量；并在“标准化残差图”中选择“直方图”和“正态概率图”。

单击“保存”按钮，打开“线性回归：保存”对话框。在“预测值”选项组中选择“未标准化”和“均值预测值的 S.E.”；在“残差”选项组中选择“未标准化”；在“预测区间”选项组中选择“均值”和“单值”。

单击“选项”按钮，打开“线性回归：选项”对话框。在“步进方法标准”中选择“使用 F 的概率”；并且选中“在等式中包含常量”选项。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
REGRESSION
   /DESCRIPTIVES MEAN STDDEV CORR SIG N
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS CI R ANOVA
   /CRITERIA=PIN(.05) POUT(.10) CIN(95)
   /NOORIGIN
   /DEPENDENT y
   /METHOD=ENTER x
   /SCATTERPLOT=(y ,*ZPRED )
   /RESIDUALS HIST(ZRESID) NORM(ZRESID)
   /CASEWISE PLOT(ZRESID) ALL
   /SAVE PRED SEPRED MCIN ICIN RESID .
```

05 设置完毕后，单击“确定”按钮，执行一元线性回归操作。

得到其结果分析如下。

- 描述性统计量

表 9.13　回归分析的描述性统计量

Descriptive Statistics

	Mean	Std.Deviation	N
肺活量	2.9025	.41442	12
体重	49.3333	5.28004	12

表 9.13 是描述性统计量的结果显示。表中显示变量 y 和 x 的均值、标准差和例数。

- 相关系数

表 9.14　相关系数

Correlations

		肺 活 量	体　　重
Pearson Correlation	肺活量	1.000	.749
	体重	.749	1.000
Sig.(1-tailed)	肺活量		.003
	体重	.003	
N	肺活量	12	12
	体重	12	12

表 9.14 是相关系数的结果。从表中可以看出，Pearson 相关系数为 0.749，单尾显著性检验的概率 ρ 值为 0.003，小于 0.05，所以体重和肺活量之间具有较强的相关性。

- 引入或剔除变量表

表 9.15　引入/剔除变量表

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	体重 [a]		Enter

a．All requested variables entered

b．Dependent Variable：肺活量

表 9.15 显示回归分析的方法以及变量被剔除或引入的信息。表中显示的回归方法是用强迫引入法引入变量 x 的。对于一元线性回归问题，由于只有一个自变量，所以此表意义不大。

- 模型摘要

表 9.16　模型摘要

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std.Error of The Estimate
1	.749[a]	.562	.518	.28775

a．Predictors: (Constant)，体重

b．Dependent Variable：肺活量

表 9.16 是模型摘要。表中显示两变量的相关系数（R）为 0.749；R Square 是判定系数，（R^2）为 0.562；Aajusted R Square 是调整判定系数，为 0.518；Std.Error of the Estimate 是估计值的标准误差，为 0.28775。

- 方差分析表

表 9.17　方差分析表

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.061	1	1.061	12.817	.005[a]
	Residual	.828	10	.083		
	Total	1.889	11			

a．Predictors: (Constant)，体重

b．Dependent Variable：肺活量

注释：在表 9.17 中，Sum of Squares 为偏差平方和，df 为样本的自由度，Mean Square 为偏差均方值，F 为 *F* 检验统计量的观察值，Sig.为自由度的熵。Regression 表示回归统计，Residual 为剩余统计。

表 9.17 是回归分析的方差分析表。从表中可以看出，回归的均方为 1.061，剩余的均方为 0.083，*F* 检验统计量的观察值为 12.817，相应的概率ρ值为 0.005，小于 0.05，可以认为变量 *x* 和 *y* 之间存在线性关系。

- 回归系数

表 9.18　回归系数

Coefficien[a]ts

Mode	Unstandardized Coefficients		Standardizec Coefficients	t	Sig.	% Confidence Interval fo	
	B	Std.Error	Beta			Lower Boun	Upper Bound
1 (Constant 体重)	.000 .059	0815 .016	.749	.001 3.580	1.000 .005	-1.815 .022	1.816 .095

a．Dependent Variable：肺活量

注释：表 9.18 中，Unstandardized Coefficients 和 Standardized Coefficients 分别表示非标准线性回归系数和标准线性回归系数，B 为回归系数，Std.Error 为标准误差，Beta 为标准化回归系数，t 为回归系数 T 检验的 t 统计量观察值，Sig.为自由度的熵，Confidence Interval for Mean 为置信区间（Lower Bound 表示置信区间的下限，Upper Bound 表示置信区间的上限）。Constant 表示常数项系数。

表 9.18 给出了线性回归方程中的参数和常数项的估计值，其中常数项系数为 0.000，回归系数为 0.059，线性回归参数的标准误差为 0.016，标准化回归系数为 0.749，回归系数 T 检验的 *t* 统计量观察值为 3.580，T 检验的概率ρ值为 0.005，小于 0.05，所以可以认为回归系数有显著意义。

- 回归诊断

表 9.19　回归诊断

Casewise Diagnostics[a]

Case Number	Std. Residual	肺 活 量	Predicted Value	Residual
1	.274	2.55	2.4711	.07889
2	-.942	2.20	2.4711	-.27111
3	.151	2.75	2.7064	.04359
4	-1.065	2.40	2.7064	.04359
5	.325	2.80	2.7064	.09359
6	-.458	2.81	2.9417	-.13172
7	1.627	3.41	2.9417	.46828
8	.550	3.10	2.9417	.15828
9	1.392	3.46	3.0594	.40063
10	-.728	2.85	3.0594	-.20937
11	.305	3.50	3.4123	.08767
12	-1.433	3.00	3.4123	-.41233

a．Dependent Variable：肺活量

注释：表 9.19 中，Case Number 为个案序号，Std.Residual 为标准化残差，Predicted Value 为预测值，Residual 为残差。

表 9.19 是对全部观察单位进行回归诊断的结果显示。从表中可以看出每一例的标准化残差、因变量 *y* 的观测值和预测值以及残差。例如第 11 例的标准化残差为 0.305。

- 残差统计量

表 9.20　残差统计量

Residuals Statistics[a]

Maximu	Mean			Std.Deviation	N
Predicted Value	2.4711	3.4123	2.9025	.31060	12
Std. Predicted Value	-1.389	1.641	.000	1.000	12
Standard Error of	.084	.165	.113	.032	12
Predicted Value	2.4436	3.6138	2.9175	.33055	12
Adjusted Predicted Value	-.4123	.46828	.00000	.27436	12
Residual	-1.433	1.627	.000	.953	12
Std. Residual	-1.748	1.701	-.023	1.050	12
Stud. Residual	-.61382	.51166	-.01501	.33567	12
Deleted Residual	-1.991	1.914	-.023	1.132	12
Stud. Deleted Residual	.016	2.694	.917	1.065	12
Mahal. Distance	.002	.747	.119	.208	12
Cook's Distance	.001	.245	.083	.097	12
Centered Leverage Value					

a．Dependent Variable：肺活量

注释：表 9.20 是残差统计量。其中 Minimum 和 Maximum 分别表示统计量的最小值和最大值，Mean 为统计量的中值，Std.Deviation 为标准差，N 为个案数量。Predicted Value 为预测值，Std.Predicted Value 为标准化预测值，Standard Error of Predicted Value 为预测值标准差，Adjusted Predicted Value Residual 为预测值调整残差，Residual 为残差值，Std.Residual 为标准残差，Stud.Residual 为学生化残差，Deleted Residual 为剔除残差，Stud.Deleted Residual 为学生化剔除残差，Mahal.Distance 为 Mahal 距离，Cook ' s Distance 为 Cook 距离，Centered Leverage Value 为中值。

通过该表可以得到预测值和残差的基本统计量信息。

- 回归标准化残差的直方图

图 9.8 是回归标准化残差的直方图，正态曲线也被显示在直方图上，用以判断标准化残差是否服从正态分布。由于本例的样本数太少，所以难以做出判断。

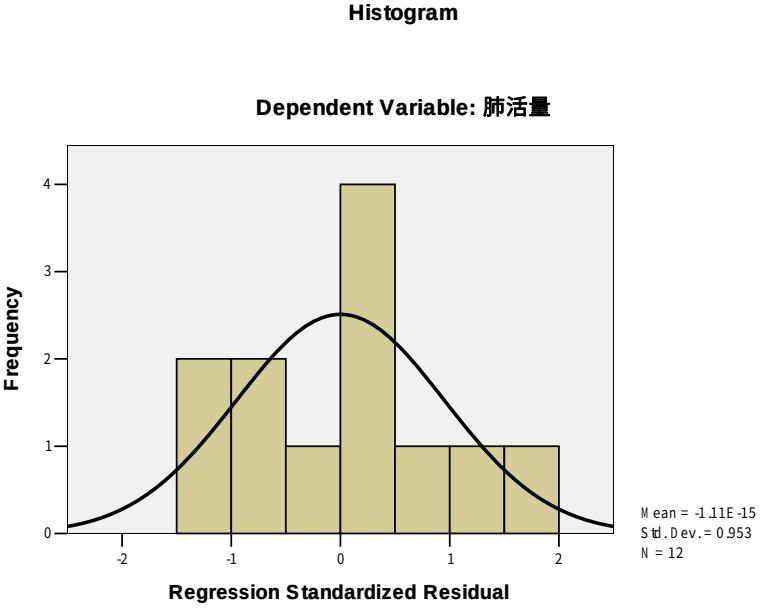


图 9.8　回归标准化残差的直方图

- 回归标准化正态 P-P 图

图 9.9 是回归标准化的正态 P-P 图。该图给出了观察值的残差分布与假设的正态分布比较，如果标准化残差呈正态分布，则标准化的残差散点应分布在直线上或靠近直线。

- 因变量与回归标准化预测值的散点图

图 9.10 显示的是因变量与回归标准化预测值的散点图，其中 DEPENDENT 为 *y* 轴变量，*ZPRED 为 *x* 轴变量。由图 9.10 可见，两变量呈直线趋势。

- 线性回归分析过程中在数据编辑窗口显示新的变量

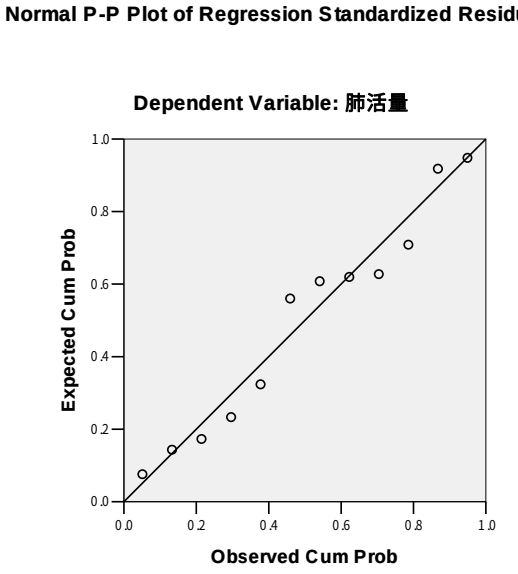


图 9.9　回归标准化正态 P-P 图

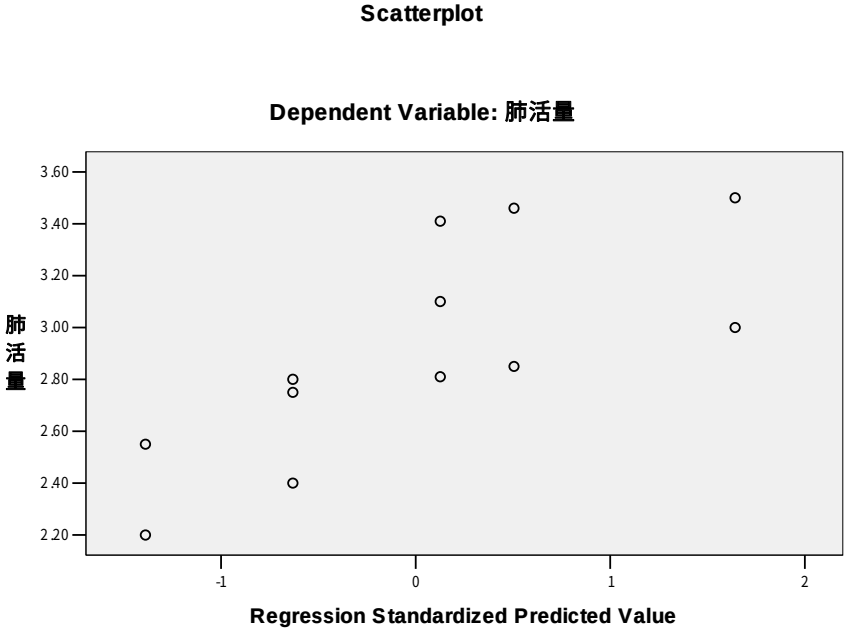


图 9.10　因变量与回归标准化预测值的散点图

图 9.11 是将 save 的结果变量加到正在使用的数据文件中，可进行线性回归的区间估计。例如，当 x=50 时预测值均值的标准误差为 0.837995，置信区间为（2.75503 3.12840）。

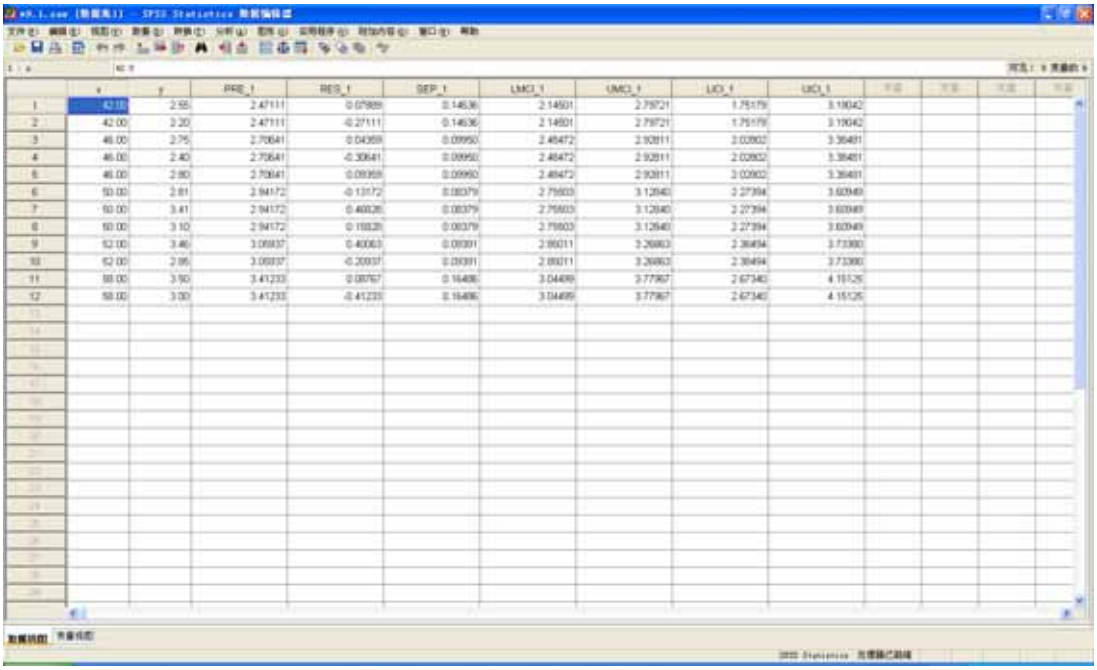


图 9.11　添加结果变量的数据编辑窗口

9.1.4　多元线性回归

【示例 9.2】小学生语言测试得分分析。

某研究机构为研究儿童的智力状况，调查了 16 所小学的平均语言测试得分（y）与家庭社会经济状况综合指标（x_1）、教师语言测试得分（x_2）及母亲教育水平（x_3）的数据，如表 9.21 所示。试进行多元线性回归分析。

表 9.21　小学的平均语言测试得分数据表

平均语言测试得分(y)	家庭社会经济状况综合指标(x_1)	教师语言测试得分(x_2)	母亲教育水平(x_3)
37.01	7.20	26.60	6.19
26.51	-11.70	24.40	5.17
36.51	12.32	25.70	7.04
40.70	14.28	25.70	7.10
37.10	6.31	25.40	6.15
41.80	12.70	24.90	6.85
33.40	-.17	25.10	5.78
44.01	9.85	26.60	6.51
23.30	-12.86	23.51	5.62
34.90	4.77	24.51	5.80

续表

平均语言测试得分(y)	家庭社会经济状况综合指标(x_1)	教师语言测试得分(x_2)	母亲教育水平(x_3)
33.10	-.96	25.80	6.19
22.70	16.04	25.20	5.62
39.70	10.62	25.10	6.94
31.80	2.60	25.01	6.33
31.70	-10.99	24.80	6.01
43.10	15.03	25.51	7.51

上述问题的 SPSS 操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开线性回归对话框和选入变量。

依次单击菜单“分析→回归分析→线性”命令，打开“线性回归”对话框（如图 9.2 所示）。在“因变量”列表框中选择变量 y，作为线性回归分析的被解释变量；在“自变量”列表框中选择 x_1、x_2、x_3，作为解释变量；在“方法”下拉列表中选择“逐步”。

03 设置。

单击“统计量”按钮，打开“线性回归：统计量”对话框（如图 9.4 所示），依次选择“描述性”、“共线性诊断”和“部分相关和偏相关性”等选项。

单击“绘制”按钮，打开“线性回归：图”对话框（如图 9.5 所示），选择“DEPENDNT”作为 y 轴变量，“*ZPRED”作为 x 轴变量；并在“标准化残差图”中选择“直方图”和“正态概率图”。

单击“保存”按钮，打开“线性回归：保存”对话框（如图 9.6 所示）。在“预测值”选项组中选择“未标准化”和“均值预测值的 S.E.”；在“残差”选项组中选择“未标准化”；在“预测区间”选项组中选择“均值”和“单值”。

单击“选项”按钮，打开“线性回归：选项”对话框（如图 9.7 所示）。在“步进方法标准”中选择“使用 F 的概率”，在“进入”文本框中输入 0.10，在“删除”文本框中输入 0.11。

其他各项采用默认选项。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
REGRESSION
    /DESCRIPTIVES MEAN STDDEV CORR SIG N
    /MISSING LISTWISE
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT y
    /METHOD=STEPWISE x1 x2 x3
    /CASEWISE PLOT(ZRESID) OUTLIERS(3) .
```

设置完毕后，单击“确定”按钮，执行多元线性回归操作。

得到其结果分析如下所示。

- 描述性统计量

表 9.22　描述性统计量

Descriptive Statistics

	Mean	Std. Deviation	N
语言测验分	34.8338	6.55982	16
家庭综合指数	4.6900	9.67199	16
教师语言测验分	25.2400	.78183	16
母亲教育水平	6.3006	.64610	16

表 9.22 是多元回归的描述性统计量表，该表显示各变量的均值、标准差和例数。

- 相关系数

表 9.23　相关系数

Correlations

		语言测验分	家庭综合指数	教师语言测验分	母亲教育水平
Pearson Correlation	语言测验分	1.000	.598	.613	.809
家庭综合指数		.598	1.000	.585	.693
教师语言测验分		.613	.585	1.000	.465
母亲教育水平		.809	.693	.465	1.000
Sig.(1-tailed)	语言测验分	.	.007	.006	.000
家庭综合指数		.007	.	.009	.001
教师语言测验分		.006	.009	.	.035
母亲教育水平		.000	.001	.035	.
N	语言测验分	16	16	16	16
家庭综合指数		16	16	16	16
教师语言测验分		16	16	16	16
母亲教育水平		16	16	16	16

表 9.23 是相关系数的结果。该表显示各变量间的 Pearson 相关系数和显著性检验单尾概率 ρ 值及例数。例如 x_3 与 y 的相关系数为 0.809，单尾概率 ρ 值为 0.000，例数为 16。

- 引入或剔除变量表

表 9.24　引入/剔除变量表

Variables Entered/Removed[a]

Mode	Variables Entered	Variables Removed	Method
1	母亲教育水平		Stepwise (Criteria Probabili y-of- F-to-enter <= .050, Probabili y-of- F-to-rem ve >= . 100)

a．Dependent Variable：语言测验分

表 9.24 显示回归分析的方法以及变量被剔除或引入的信息。从表中可以看出，模型最

先引入变量 x_3，第二个引入模型的是变量 x_2，没有变量被剔除，所有模型 2 中有变量 x_3 和 x_2。

- 模型摘要

表 9.25　模型摘要

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.809[a]	.655	.630	3.98861

a．Predictors: (Constant)，母亲教育水平

b．Dependent Variable：语言测验分

表 9.25 是模型摘要。该表显示各模型的拟合情况。从表中可以看出，模型 2 的复相关系数为 0.809，判定系数为 0.655，调整判定系数为 0.630（若将作用不显著的变量引入方程，则该系数减少），估计值的标准误差为 3.98861。

- 方差分析表

表 9.26　方差分析表

ANOV[b]

Mode		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	422.742	1	422.742	26.572	.000[a]
	Residual	222.727	14	15.909		
	Total	645.468	15			

a．Predictors:(Constant)，母亲教育水平

b．Dependent Variable：语言测验分

表 9.26 是回归分析的方差分析表，该表显示一个模型的方差分析结果。从表中可以看出，模型 Regression 的 F 统计量的观察值为 26.572，概率 ρ 值为 0.000，在显著性水平为 0.05 的情形下，可以认为 y（语言测试得分）与 x_3（母亲教育水平）和 x_2（教师语言测验得分）之间有线性关系。

- 回归系数

表 9.27　回归系数

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-16.936	10.092	.809	-1.678	.116
	母亲教育水平	8.217	1.594		5.155	.000

a．Dependent Variable：语言测验分

表 9.27 为线性回归系数列表，表中显示各模型的偏回归系数、标准误差、常数、标准化偏回归系数、回归系数检验的 t 统计量观测值和相应的概率 ρ 值。根据模型 2 建立的多元线性回归方程为：

$$y = 8.217 + 1.594x_2 + 0.809x_3$$

方程中的常数项为 8.217，偏回归系数 b_2 为 1.594，b_3 为 0.809，经 T 检验，b_2 和 b_3 的概率 ρ 值分别为 0.116 和 0.000，按给定的显著性水平情形下，均有显著性意义。

- 模型外的变量

表 9.28 显示回归方程外的各模型变量的有关统计量，即 Beta t 值和概率 ρ 值、偏相关性和共线性的容许值，可见模型 Regression 方程外的各变量偏回归系数经重检验，概率 ρ 值均大于 0.10，故不能引入方程。

表 9.28　模型外的变量

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	家庭综合指数	.072[a]	.319	.755	.088	.520
	教师语言测验分	.302[a]	1.844	.088	.455	.784

a．Predictors in the Model: (Constant),，母亲教育水平

b．Dependent Variable：语言测验分

表 9.28 显示的是回归方程外的各模型变量的有关统计量，即 Beta t 值和概率 ρ 值、偏相关性和共线性的容许值，可见模型 2 方程外的各变量偏回归系数经重检验，概率 ρ 值分别为 0.520 和 0.784，均大于给定的显著性水平，故不能引入方程。

- 残差统计量

表 9.29　残差统计量

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	25.5439	44.7707	34.8338	5.30874	16
Residual	-6.54133	7.45590	.00000	3.85337	16
Std. Predicted Value	-1.750	1.872	.000	1.000	16
Std. Residual	-1.640	1.869	.000	.966	16

a．Dependent Variable：语言测验分

表 9.29 是回归模型的残差统计量，标准化残差的绝对值最大为 1.869，没有超过默认值 3，则显示具体观察单位的标准化残差，以发现奇异值。

9.2　非线性回归

非线性回归是指在因变量与一系列自变量之间建立非线性模型。“ 线性 ”和“ 非线性 ”并不是说因变量和自变量之间是直线或曲线关系，而是说因变量是否能用自变量的线性组合来表示。如果经过变量转换，两个变量可以用线性鞭打其关系，那么可以用上一节介绍的方法；如果经过变量变换后，两个变量关系仍然不能用线性形式表达，就可以用本节介绍的非线性回归分析。

例如，血药浓度—时间曲线是先升后降；药剂量和疗效反应率之间的关系呈曲线变化趋势。有时，在局部内两个变量的关系也许呈直线趋势，扩大范围后却显示出曲线趋势。如人的生长发育，在某一阶段，身高与年龄呈直线关系，但是从整个生命周期看，身高与年龄之间却是明显的曲线关系。

9.2.1　非线性回归的基本原理

一般非线性回归模型可以表示为：

$$\mu_{y|x} = f(\beta_1, \beta_2, ..., \beta_p, x) \tag{9.10}$$

其中，x 为自变量，可以是一个，也可以是多个；$\beta_1, \beta_2, ..., \beta_p$ 为总体回归系数；y 是关注的因变量，$\mu_{y|x}$ 为给定 x 时 y 的总体均值。模型中除了自变量和因变量的关系为非线性外，其他假定条件与线性回归基本上相同。

非线性回归是通过迭代算法实现的。SPSS 采用的迭代算法有两种：Levenberg-Marquardt 法和序列二次规划法。

Levenberg-Marquardt 法又叫阻尼最小二乘法，是 Gauss-Newton 法的改进算法。它有一个阻尼因子 λ，用 λ 可以控制搜索步长和方向。当 λ=0 时，即为 Gauss-Newton 法；当 $\lambda \to \infty$ 时，趋于零向量，即为最快下降法。Levenberg-Marquardt 法的优势在于对影响 Gauss-Newton 法有效性的病态二次项，也可以通过阻尼因子 λ 来控制。

序列二次规划法的主要思路是：形成基于拉格朗日函数二次近似的二次规划子问题，而这些问题可以用任意一种二次规划算法求解，求得的解用来形成新的迭代公式，作为下一次搜索的依据。用序列二次规划算法求解非线性有约束问题时的迭代次数常常比求解无约束问题时少，因为在搜索区内，序列二次规划算法可以获得最佳的搜索步长和方向信息。

9.2.2　SPSS 非线性回归

非线性回归分析中因变量和自变量要求是定量变量，如果自变量是分类变量，则应先转换为二分类的哑变量。

SPSS 非线性回归的操作步骤如下。

01 选择非线性回归对话框。

依次单击菜单“ 分析→回归分析→非线性 ”命令，打开“ 非线性回归 ”对话框，如图 9.12 所示。

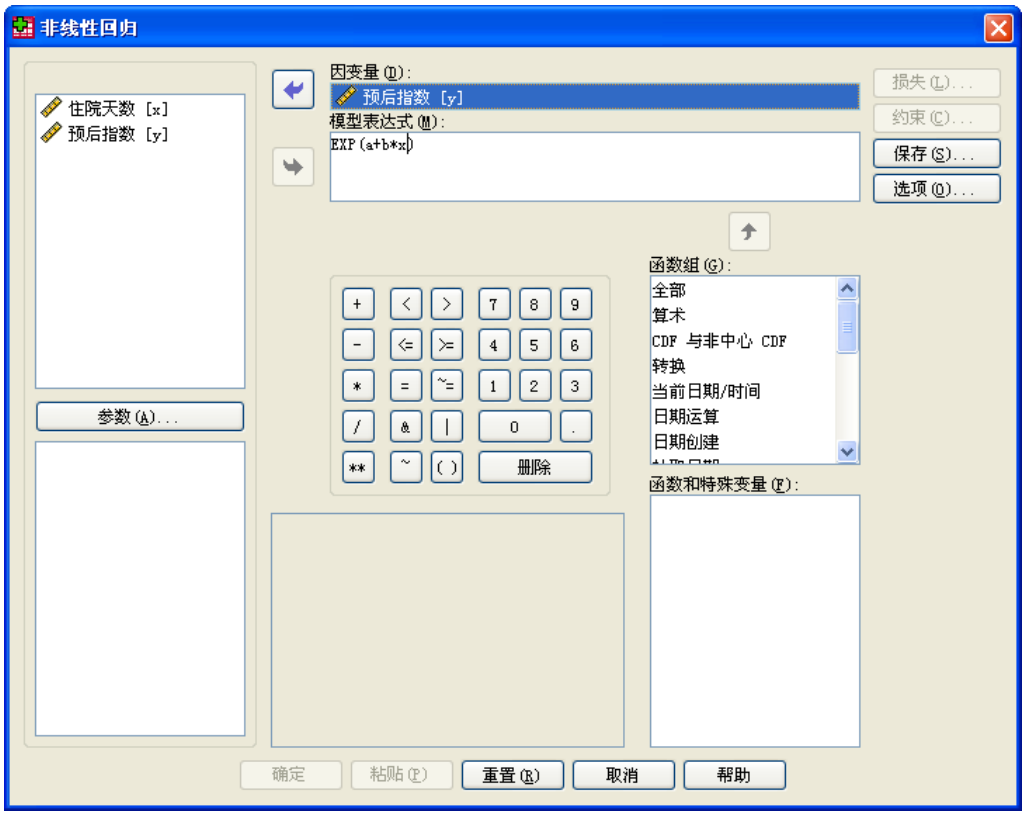


图 9.12 “ 非线性回归 ”对话框

用户可以将左侧列表框中的变量选入到“ 因变量 ”列表框中。因变量可以是数值型的，也可以是分类变量，如果是分类变量，则在非线性回归分析前应进行转换。用户可以在“ 模型表达式 ”中输入模型，输入的模型至少应包含一个自变量。在“ 函数组 ”列表框中给出了各种可能用到的数学函数，用户可以选择使用。

02 设置。

单击图 9.12 中的“ 参数 ”按钮，打开“ 非线性回归：参数 ”对话框。如图 9.13 所示。在进行迭代计算来确定模型参数时，首先必须给定参数的初值。在“ 非线性回归：参数 ”对话框内可以指定模型参数的初值。“ 名称 ”文本框指定参数的名称，必须是合法的，并且是模型表达式中使用的名称；“ 初始值 ”文本框指定参数的初值，初值越接近最终确定的参数真值越好，所有参数都需要指定初值，不合适的初值会导致迭代不收敛或建立的模型只对部分数据有效，将前次计算的参数结果作为当前初值，可以增加计算的精度。指定完“ 名称 ”和“ 初始值 ”后，单击“ 添加 ”按钮可以将其加入到右侧的列表框中。如果选

择“ 使用上一分析的起始值 ”，它将取代事先指定的初值，该选项在后面的分析中一直起作用，所以当它变换模型时，务必不要忘记取消该选项。

单击图 9.12 对话框中“ 损失 ”按钮，打开“ 非线性回归：损失函数 ”对话框，如图 9.14 所示。损失函数，是指非线性回归中通过运算使最小化的函数，必要时损失函数可以分段表示。用户可以选择是以“ 残差平方和 ”为损失函数还是“ 用户定义的损失函数 ”。如果用户选择“ 用户定义的损失函数 ”，可以从左侧的列表框选择。如 RESID_**2，表示的就是最小二乘法。



图 9.13 “ 非线性回归：参数 ”对话框



图 9.14 “ 非线性回归：损失函数 ”对话框

单击图 9.12 对话框中 “ 约束 ”按钮，打开“ 非线性回归：参数约束 ”对话框。如图 9.15 所示。用于设置参数约束，是针对在得到最终参数值的迭代过程中所允许参数的取值范围而言的。用户可以选择“ 未约束 ”或“ 定义参数约束 ”。



图 9.15 “ 非线性回归：参数约束 ” 对话框

单击图 9.12 对话框中“ 保存 ” 按钮，打开“ 非线性回归：保存新变量 ” 对话框。如图 9.16 所示。用于设置需要保存的统计量，可以是“ 预测值 ”、“ 残差 ”、“ 导数 ” 和“ 损失函数值 ”。这些统计量在分析模型的拟合程度或检查可能有问题的观察值时很有用。

单击图 9.12 对话框中“ 选项 ” 按钮，打开“ 非线性回归：选项 ” 对话框，如图 9.17 所示。用于设置与分析方法有关的选项。

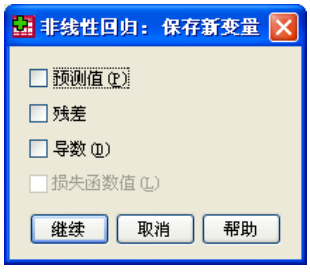


图 9.16 “ 非线性回归：保存新变量 ” 对话框

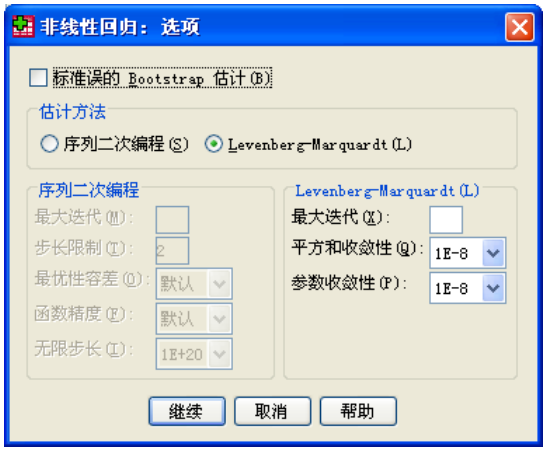


图 9.17 “ 非线性回归：选项 ” 对话框

03 执行操作。

设置完毕后，单击图 9.12 对话框中“ 确定 ” 按钮，执行非线性回归。

【示例 9.3】非线性回归

一位医院管理人员想建立一个回归模型，对重伤病人出院后的长期恢复情况进行预

测。自变量为病人住院天数（x），因变量为病人出院后长期恢复的预后指数（y），指数取值越大表示预后结果越好。数据如表 9.30 所示。

表 9.30　15 名重伤病人的住院天数 x（天）与预后指数 y

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
住院天数 x	2	5	7	10	14	19	26	31	34	38	45	52	53	60	65
预后指数 y	54	50	45	34	35	25	20	16	18	13	8	11	8	4	6

首先，绘制散点图，如图 9.18 所示。

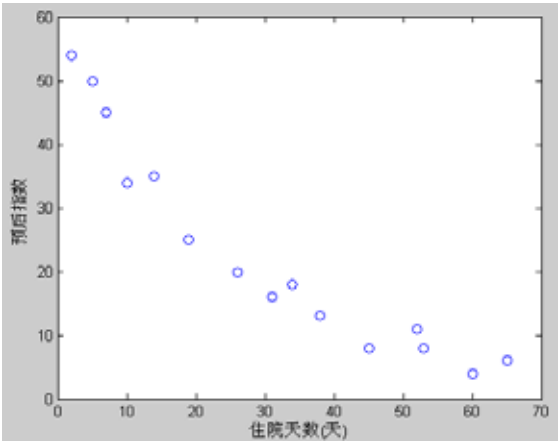


图 9.18　15 名重伤病人的住院天数与预后指数的散点图

对两个变量可尝试拟合指数曲线如 $\overline{y}=e^{(b_0+b_1x)}$，对因变量 y 做自然对数变换，得到：$y'=\ln y$。观察 y' 与 x 的散点图，y' 与 x 呈直线趋势（如图 9.19 所示）。注意，如果此时用最小二乘法拟合 y' 与 x 的直线回归方程 $\hat{y}'=b_0+b_1x$，之后再将其结果带回 $\hat{y}=e^{\hat{y}'}$，那么得到的方程不能保证残差平方和 $\sum(y-\hat{y})^2$ 最小，因为此时方程 $\hat{y}'=b_0+b_1x$ 只保证了 $\sum(y'-\hat{y}')^2$ 最小。非线性回归中迭代算法得到的方程 $\hat{y}=e^{(b_0+b_1x)}$，可以保证残差平方和 $\sum(y-\hat{y})^2$ 最小。

上述问题的 SPSS 操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开非线性回归对话框并选入变量。

依次单击菜单“ 分析→回归分析→非线性 ”命令，打开“ 非线性回归 ”对话框。在“ 因变量 ”列表框中选择变量“ 预后指数 ”;在“ 模型表达式 ”中输入“ EXP(a+b*住院天数（x）) ”。

单击“ 参数 ”按钮，在“ 名称 ”文本框中输入“ a ”,“ 初始值 ”文本框中输入“ 4 ”，单击“ 添加 ”按钮；在“ 名称 ”文本框中输入“ b ”,“ 初始值 ”文本框中输入“ -0.04 ”，单击“ 添加 ”按钮。参数初始值“ 4 ”与“ –0.04 ”是根据曲线直线化后的最小二乘拟合的模

型参数进行估计的。

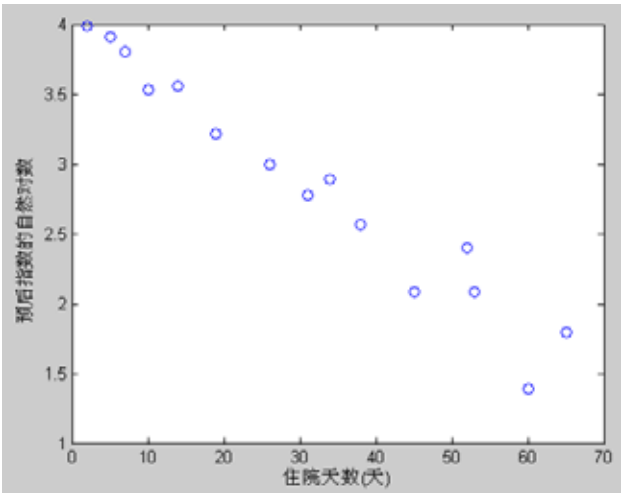


图 9.19　预后指数的自然对数与住院天数的散点图

其他各项采用默认选项。

03 单击“粘贴”按钮得到的 Syntax 命令语句程序如下：

```
* NonLinear Regression.
    MODEL PROGRAM a=4 b=-0.04 .
    COMPUTE PRED_ = EXP(a+b*住院天数（x）).
    NLR 预后指数（y）
    /OUTFILE='C:\WINDOWS\TEMP\spss2420\SPSSFNLR.TMP'
    /PRED PRED_
    /CRITERIA SSCONVERGENCE 1E-8 PCON 1E-8 .
```

设置完毕后，单击“确定”按钮，执行曲线估计操作。

得到其结果，分析如下。

- 各次迭代步骤中的残差、参数计算值

表 9.31　各次迭代步骤中的残差、参数计算值

Iteration History[b]

Iteration Number[a]	Residual Sum of Squares	Parameter	
		a	b
1.0	112.821	4.000	-.040
1.1	49.562	4.074	-.040
2.0	49.562	4.074	-.040
2.1	49.459	4.071	-.040
3.0	49.459	4.071	-.040
3.1	49.459	4.071	-.040
4.0	49.459	4.071	-.040

续表

Iteration Number[a]	Residual Sum of Squares	Parameter	
		a	b
4.1	49.459	4.071	-.040

Derivatives are calculated numerically.

a．Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b．Run stopped after 8 model evaluations and 4 derivative evaluations because the relative reduction between successive residual sums of squares is at most SSCON = 1.00E-008.

注释：表 9.31 中，Iteration Number 为重复次数，Residual Sum of Squares 为残差平方和，Parameter 为参数计算值，其中 a 为系数项，b 为指数项。

表 9.31 给出了每一次迭代步骤中各次的残差、参数计算值。迭代经过 8 次模型计算和 4 次求导计算后终止，两次相邻计算的残差平方和的差值几乎等于 1.00E-008。

- 参数估计值

表 9.32　参数估计值

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	4.071	.025	4.017	4.125
b	-.040	.002	-.043	-.036

表 9.32 给出了参数估计值、渐近标准差和渐近 95%置信区间。参数 a 的估计值为 4.071，参数 b 的估计值为–0.040。两者的 95%的置信区间均不包括 0，表明参数 a 和 b 均有统计学意义。

- 参数相关系数

表 9.33　参数 a 和 b 相关系数信息

Correlations of Parameter Estimates

	a	b
a	1.000	-.707
b	-.707	1.000

表 9.33 给出了参数 a 和 b 相关系数，为–0.707。

- 非线性回归模型检验结果

表 9.34　非线性回归模型检验结果

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	12060.541	2	6030.270

续表

Source	Sum of Squares	df	Mean Squares
Residual	49.459	13	3.805
Uncorrected Total	12110.000	15	
Corrected Total	3943.333	14	

Dependent variable：预后指数（y）

a．R squared = 1 -（Residual Sum of Squares）/

(Corrected Sum of Squares) = .987.

表 9.34 给出了非线性回归模型的检验结果，包括回归项、残差项、没有校正和校正后总的自由度、平方和与均方的大小。决定系数为 0.987，表明所得回归模型拟合效果很好。

建立的回归方程为：

$$\hat{y} = e^{(4.071-0.043x)}$$

表 9.35 给出了进行曲线直线化后回归和非线性回归的结果比较。

模型名称	R^2	a	b
曲线直线化后回归	0.955	4.037	-0.038
非线性回归	0.987	4.071	-0.043

从表 9.35 可以看出，采用非线性回归所得的结果比直线化后线性回归所得的结果有所改善。

9.3 曲线估计

对两个变量间不呈直线关系的资料，我们可以直接进行曲线拟合。曲线拟合（Curve Fitting）是求解反映变量曲线关系的曲线估计方程（Curvilinear Regression Equation）的过程。

曲线拟合的一般步骤如下。

01 根据自变量 x 和因变量 y 散点图呈现的趋势，结合专业知识及经验选择合适的曲线形式。在某些情况下，绘制散点图时采用一些特殊的坐标系可能更有利于揭示变量间的关系，更容易确定曲线方程的形式。例如，在半对数坐标系中，散点呈现较为明显的直线趋势，即可选用指数曲线 $\hat{y} = e^{(b_0+b_1x)}$ 或对数曲线 $\hat{y} = b_0 + b_1 \ln x$。

02 选用适当的估计方法求得回归方程。如果曲线形式可表示为 x 的某种变化形式与 y 的线性关系（六人，对数曲线 $\hat{y} = b_0 + b_1 \ln x$），即可采用“曲线直线化”的方法对变换后的 z（如 $z = \ln x$）和 y 做最小二乘拟合；如果曲线形式表示为 y 的某种变换形式 $\hat{y}^*$ 与 x 的线性关系（例如，将指数曲线 $\hat{y} = e^{(b_0+b_1x)}$ 变换为 $\hat{y}^* = b_0 + b_1x$），则可采用“非线性最小二乘”估计方法。

03 在实际工作中，有时可结合散点图试拟合几种不同形式的曲线方程并计算 R^2，一般来说， R^2 较大时拟合效果较好。但应注意，为了单纯地得到较大的 R^2，模型的形式可

能会很复杂，甚至使其中的参数无法解释实际意义，这是不可取的。因此，要充分考虑专业知识，结合实际解释和应用效果来确定最终的曲线形式。

决定系数 R^2 定义为：

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}=1-\frac{SS_{残}}{SS_{总}} \tag{9.11}$$

曲线估计是 Regression 的一个内容，它可以用于拟合许多常用的曲线，理论上只要两个变量间存在某种可以被它描述的数量关系，就可以用该过程来分析处理。

下面介绍曲线估计过程会使用到的界面、对话框和选项。

- 选择曲线估计对话框

依次单击菜单“ 分析→回归分析→曲线估计 ”命令，打开“ 曲线估计 ”对话框，如图 9.20 所示。

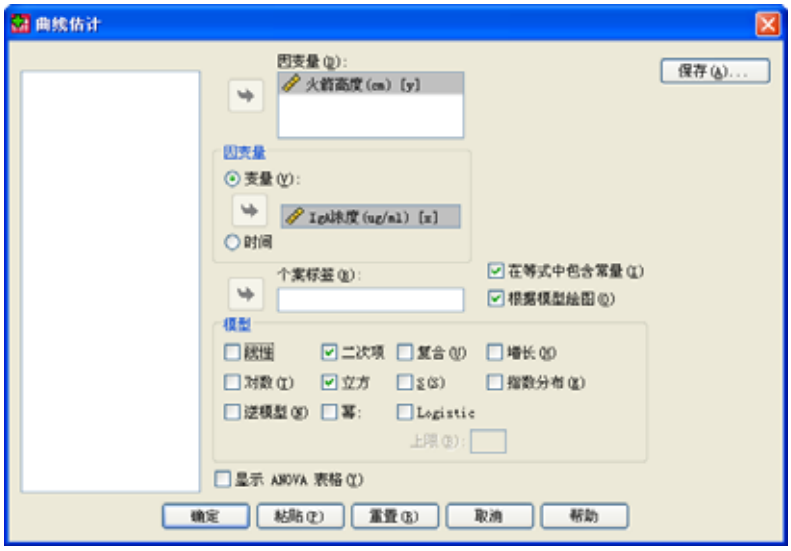


图 9.20 “ 曲线估计 ”对话框

左侧框内包含数据所有变量名。用户可以选择变量，单击 按钮，将其选入到右侧的“ 因变量 ”窗口，可以选入多个，如果选入多个则对各个因变量分别拟合模型；还可以将选择的变量作为时间自变量或普通自变量选入“ 自变量 ”列表框。

在“ 模型 ”选项组中列出了曲线拟合的模型。根据两个变量散点图显示的曲线趋势选择合适的拟合模型，是该对话框的重点。模型中包含的选项及其含义如表 9.36 所示。

表 9.36　模型选项及其含义

选　项	含　义
线性	拟合直线方程，与“ 线性拟合 ”过程的直线回归相同
对数	拟合对数方程 $\hat{y}=b_0+b_1\ln x$

续表

选　项	含　义
逆模型	拟合方程 $\hat{y}=b_0+b_1/x$
平方	拟合二次方程 $\hat{y}=b_0+b_1x+b_2x^2$
立方	拟合三次方程 $\hat{y}=b_0+b_1x+b_2x^2+b_3x^3$
幂	拟合乘幂曲线模型 $\hat{y}=b_0x^{b_1}$
混合	拟合复合曲线模型 $\hat{y}=b_0b_1^x$
S	拟合 S 形曲线 $\hat{y}=e^{(b_0+b_1/x)}$
logistic	拟合逻辑曲线模型 $\hat{y}=1/(1/u+b_0b_1^x)$。选择此模型，“上限”文本框被激活，用户可以输入数值，作为上限
增长	拟合复合比级数曲线模型 $\hat{y}=e^{(b_0+b_1x)}$
指数分布	拟合指数方程 $\hat{y}=b_0e^{b_1x}$

“在等式中包含常量”可以在方程中包含常数项，为对话框默认选项。

“根据模型绘图”包括原始数值的连线图和拟合模型的曲线图，它在曲线拟合中是非常重要的，也是默认选项。

用户可以选中“显示 ANOVA 表格”选项来显示方差分析表。

- “保存”选项

在图 9.20 所示对话框中单击“保存”按钮，打开“曲线估计：保存”对话框。如图 9.21 所示。

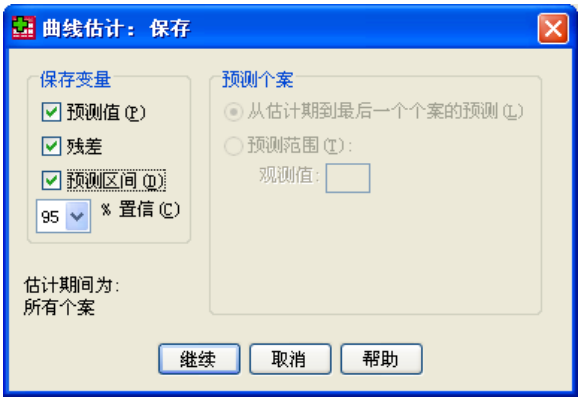


图 9.21 “曲线估计：保存”对话框

该对话框用于设置存储中间结果，如预测值、残差、预测区间等。

【示例 9.4】曲线估计示例。

用已知浓度的免疫球蛋白 A（IgA，$\mu g/ml$）做火箭电泳，测得火箭高度（cm），如表 9.37 所示。试采用恰当的回归方程描述火箭高度 y 与 IgA 浓度 x 之间的关系。

表 9.37　火箭高度 *y* 与 IgA 浓度 *x* 数据

样品编号	1	2	3	4	5	6	7	8
IgA 浓度 *x*	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
火箭高度 *y*	7.6	12.3	15.7	18.2	18.7	21.4	22.6	23.8

首先对火箭高度和免疫球蛋白 A 浓度绘制散点图（如图 9.22 所示）。

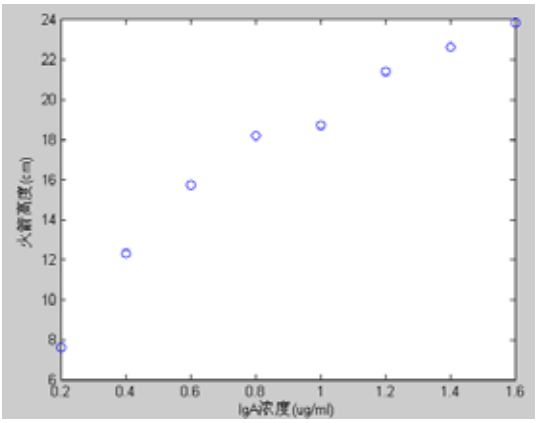


图 9.22　火箭高度与免疫球蛋白 A 浓度的散点图

从图 9.22 可以看出，两者的斜率有逐渐减缓的趋势，这里选用二次曲线模型、三次曲线模型和对数模型。拟合三个模型，将三者拟合情况进行比较，选择拟合优度好的模型。

上述问题的 SPSS 操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开曲线估计对话框和选入变量。

依次单击菜单“ 分析→回归分析→曲线估计 ”命令，打开“ 曲线估计 ”对话框（如图 9.20 所示）。在“ 因变量 ”列表框中选择变量“ 火箭高度 *y* ”；在“ 自变量 ”列表框中选择“ IgA 浓度 *x* ”；在“ 模型 ”选项组中选择“ 平方 ”、“ 立方 ”和“ logistic ”。

其他各项采用默认选项。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
* Curve Estimation.
CURVEFIT /VARIABLES=y  WITH x
    /CONSTANT
    /MODEL=QUADRATIC CUBIC LGSTIC
    /PLOT FIT
    /CIN=95
    /SAVE=PRED RESID CIN.
```

设置完毕后，单击“ 确定 ”按钮，执行曲线估计操作。

其结果分析如下。

- 模型拟合描述

表 9.38　模型拟合描述

Model Description

Model Name		MOD_1
Dependent Variable	1	火箭高度（cm）
Equation	1	Quadratic
	2	Cubic
	3	Logistic
Independent Variable		IgA 浓度（ug/ml）
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified
Tolerance for Entering Terms in Equations		.0001

a．The model requlres all non-missing values to be positive.

注释：表 9.38 中，Model Name 为模型的名称，Dependent variable 为因变量，Equation 为回归方程式，Independent Variable 为自变量，Constant 为常数项，Variable Whose Values Label Observation in Plots 为变量在图形中的观测值，Tolerance for Entering Terms in Equations 为对回归方程进行插值处理的容许量。

表 9.38 是对模型拟合过程做的一些描述，给出因变量数量和变量名、拟合模型的数量和类型、自变量变量名、回归方程包括常数项等情况。

- 拟合样本例数信息

表 9.39　拟合样本例数信息

Case Processing Summary

	N
Total Cases	8
Excluded Cases[a]	0
Forecasted Cases	0
Newly Created Cases	0

a．Cases with a missing value in any variable are excluded from the analysis.

注释：表 9.39 中，Total Cases 为个案总数，Excluded Casesa 为被拒绝的个案数，Forecasted Cases 为预测的个案数，Newly Created Cases 为新建个案数。

表 9.39 是对进行拟合的样本例数进行说明的信息。

- 变量拟合信息

表 9.40　变量拟合信息

Variable Processing Summary

	Variables	
	Dependent	Independent
	火筒高度（cm）	IgA 浓度（ug/ml）
Number of Positive Values	8	8
Number of Zeros	0	0
Number of Negative Values	0	0
Number of Missing Values User-Missing	0	0
System-Missing	0	0

注释：表 9.40 中，Dependent 为因变量，Independent 为自变量。Number of Positive Values 为正值个案数，Number of Zeros 为零值个案数，Number of Negative Values 为负值个案数，Number of Missing Value 为丢失值个案数（User-Missing 表示用户丢失，System-Missing 表示系统丢失）。

表 9.40 给出了变量拟合过程的一些情况。

- 拟合的三个回归模型的检验报告

表 9.41　拟合的三个回归模型的检验报告

Model Summary and Parameter Estimates

Dependent Variable：火筒高度(cm)

Equation	Model Summary						Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Quadratic	.987	185.165	2	5	.000	4.091	21.872	-6.116	6.471
Cubic	.994	229.287	3	4	.000	1.529	35.202	-23.588	
Logistic	.842	32.088	1	6	.001	.115	.490		

The independent variable is IgA 浓度(ug/ml).

注释：表 9.41 中，Equation 为回归方程式，Model Summary 为模型摘要信息（R Sequare 为判定系数，F 为检验统计量的观察值，df1 和 df2 为自由度，Sig.为自由度的熵），Parameter Estimates 为模型估计参数（Constant 表示常数项系数，b1、b2 和 b3 分别为 1 次项、2 次项和 3 次项系数）。Quadratic、Cubic 和 Logistic 分别表示三种回归模型。

表 9.41 给出了所拟合的三种回归模型的检验报告，包括拟合优度、模型检验结果和各个参数值。结果显示，三种回归模型均有统计学意义。由拟合优度来确定最佳的模型，三次方曲线的拟合优度最好，应选择该模型，但是三次方曲线的参数比较多，相对来说较为复杂。而对数曲线模型的优度也很不错，和三次方曲线的拟合优度相差很小，因此选择对数曲线模型。

- 不同模型的拟合结果图

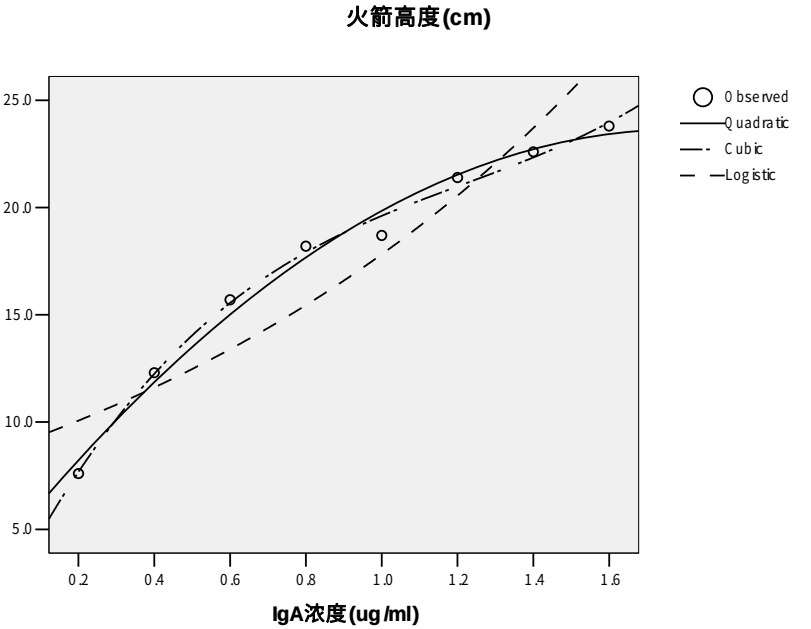


图 9.23　不同模型的拟合结果图

图 9.23 是三个模型的曲线和实际测量值的连线情况，对数曲线和三次方曲线对模型拟合相差很小，只是在浓度小于 $0.2\,\mu g/ml$ 时，三次方曲线稍优于对数曲线。在曲线估计中，模型的简洁和拟合优度好坏一样重要，因此选择对数曲线模型。

建立的回归方程为：

$$\hat{y} = 19.745 + 7.777 \ln x$$

表 9.42 给出了该资料数据的直线回归模型和对数曲线模型的结果比较，可见两个模型的拟合结果是相同的。

表 9.42　两种方法拟合回归模型的结果比较

模型名称	R^2 值	F 值	P 值
曲线直线化	0.992	763.499	0.000
对数曲线	0.992	763.499	0.000

9.4　logistic 回归

在前面介绍的回归模型中，因变量为区间（定量）变量，并且理论上要求其服从正态分布等 LINE（线性、独立、正态、等方差）假定条件。本节所介绍的 logistic 回归与前面介绍的回归方法非常类似，它们之间的主要区别在于：因变量的类型不同。通过一组预报变量（即一组自变量，也称为解释变量或协变量），采用 logistic 回归，可以预测一个分

类变量每一分类所发生的概率。因变量为分类变量，预报变量可以是区间变量，也可以是分类变量，还可以是区间二与分类变量的混合。如果自变量均为区间变量，则这类数据也可采用判别分析等方法进行分析，但通常情况下，logistic 回归对预报变量（自变量）的假定条件较少，所以 logistic 回归更为常用。

分类变量分为有序分类变量（即有序多项分类变量）和无序分类变量两种；而无序分类变量也叫名义变量，可分为二项分类变量和无序多项分类变量两种。在实际工作中，因变量为分类变量的例子很多，例如，经某种方案处理后，病人的治疗效果分为生存和死亡，有效与无效（二项分类）；本科生经 4~5 年大学学习后，对大学生活的满意程度分为很不满意、不满意、满意、很满意，结果变量满意程度为有序分类变量；不同人群将会选择不同品牌（如佳能、柯达、富士、索尼）的数码相机，这里的结果变量相机品牌为无序多项分类变量。

下面介绍 logistic 回归过程会使用到的界面、对话框和选项。

- 选择 logistic 回归对话框

依次单击菜单“ 分析→回归分析→二元 Logistic 回归 ”命令，打开“ Logistic 回归 ”对话框，如图 9.24 所示。

- 选择变量

在变量列表框中选择变量名以后，单击 按钮，将其移动到“ 因变量 ”列表框中；然后选择一个或多个变量，单击 按钮，将其移动到“ 协变量 ”列表框。



图 9.24 “ Logistic 回归 ”对话框

- 设置

在图 9.24 所示的“ Logistic 回归 ”对话框中单击“ 方法 ”下拉菜单项可以选择自变量进入模型的多种方法，如表 9.43 所示。

表 9.43　方法选项及其含义

方 法 名	含　义
进入	强迫引入法，这是 SPSS 的默认选项，即将所选自变量全部放入模型中
向前：条件	（条件似然比）向前逐步法
向前：LR	（似然比）向前逐步法
向前：Wald	（Wald）向前逐步法
向后：条件	（条件似然比）向后逐步法
向后：LR	（似然比）向后逐步法
向后：Wald	（Wald）向后逐步法

用户还可以将一个变量选入“选择变量”列表框，根据该变量的值，通过右侧的“规则”按钮，可以建立一个选择条件，可以只对部分数据进行分析。

- “分类”设置

单击“分类”按钮，打开“Logistic回归：定义分类变量”对话框，如图9.25所示。



图 9.25 “Logistic 回归：定义类型变量”对话框

当变量不是连续型变量，而是分类变量时，采用此对话框，计算机可自动对这类变量进行变量类型化。在输出结果中，会提示所选每一分类变量的具体编码情况，解释结果时应特别注意这些信息，因为不同的编码方法将会得到不同的回归系数。

- “保存”设置

单击“保存”按钮，打开“Logistic 回归：保存”对话框，如图 9.26 所示。

在 logistic 回归分析中，有很多与每一个观察个体有关的重要信息，可以通过这个对话框保存下来，如预测值、残差等。

- “选项”设置

单击“选项”按钮，打开“Logistic 回归：选项”对话框，如图 9.27 所示。

通过这个对话框，可获得 Hosmer-Lemshow 拟合度检验结果和预测概率分类图。

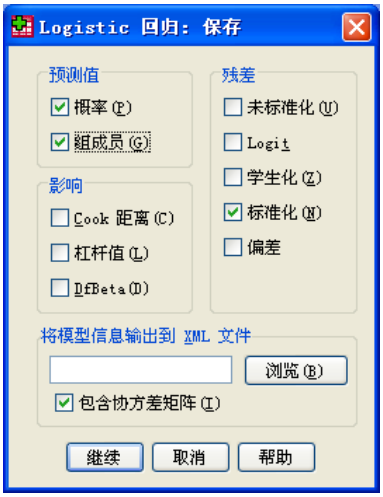


图 9.26 “ Logistic 回归：保存 ”对话框

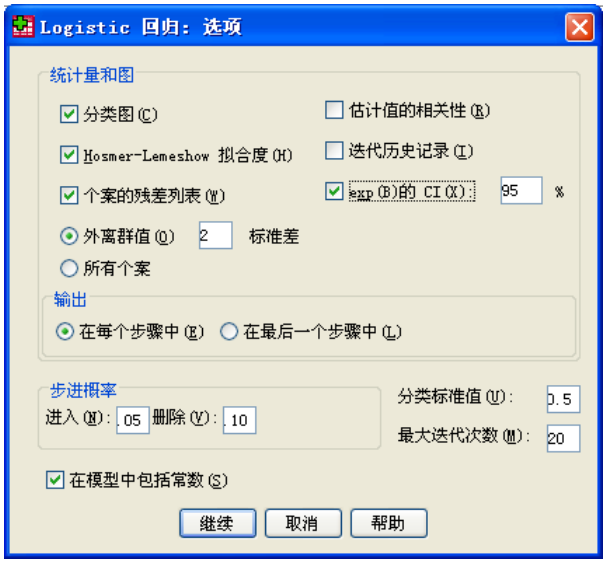


图 9.27 “ Logistic 回归：选项 ”对话框

有关 Logistic 回归对话框中的具体设置及其含义与前面的分析都大致相同，在此不再赘述。

【示例 9.5】logistic 回归示例。

前列腺癌细胞是否扩散到邻近的淋巴结，是选择治疗方案的重要依据。为了了解淋巴组织中有无癌细胞转移，通常的做法是对病人实施剖腹术探查，并在显微镜下检查淋巴组织。为了不手术而又能弄清淋巴结的转移情况，Brown(1980 年)在术前检查了 53 例前列腺癌患者，分别记录了年龄（AGE）、酸性磷酸酯酶（ACID）两个连续型变量和 X 射线（X-RAY）、术前探针活检病理分基（GRADE）、直肠指检肿瘤的大小与位置（STAGE）三个分类变量。后三个变量均按 0、1 赋值，其值 1 表示阳性或较严重情况，0 表示阴性或较轻情况。还有手术探查结果变量 NODES，1 表示有淋巴结转移，0 表示无淋巴结转移，数据如表 9.44 所示。

表 9.44　53 例接受手术的前列腺癌患者淋巴结转移情况

No.	X_RAY	GRADE	STAGE	AGE	ACID	NODES
1	0	1	1	64	40	0
2	0	0	1	63	40	0
3	1	0	0	65	46	0
4	0	1	0	67	47	0
5	0	0	0	66	48	0
6	0	1	1	65	48	0
7	0	0	0	60	49	0
8	0	0	0	51	49	0
9	0	0	0	66	50	0

续表

No.	X_RAY	GRADE	STAGE	AGE	ACID	NODES
10	0	0	0	58	50	0
11	0	1	0	56	50	0
12	0	0	1	61	50	0
13	0	1	1	64	50	0
14	0	0	0	56	52	0
15	0	0	0	67	52	0
16	1	0	0	49	55	0
17	0	1	1	52	55	0
18	0	0	0	68	56	0
19	0	1	1	66	59	0
20	1	0	0	60	62	0
21	0	0	0	61	62	0
22	1	1	1	59	63	0
23	0	0	0	51	65	0
24	0	1	1	53	66	0
25	0	0	0	58	71	0
26	0	0	0	63	75	0
27	0	0	1	53	76	0
28	0	0	0	60	78	0
29	0	0	0	52	83	0
30	0	0	1	67	95	0
31	0	0	0	56	98	0
32	0	0	1	61	102	0
33	0	0	0	64	187	0
34	1	0	1	58	48	1
35	0	0	1	65	49	1
36	1	1	1	57	51	1
37	0	1	0	50	56	1
38	1	1	0	67	67	1
39	0	0	1	67	67	1
40	0	1	1	57	67	1
41	0	1	1	45	70	1
42	0	0	1	46	70	1
43	1	0	1	51	72	1
44	1	1	1	60	76	1
45	1	1	1	56	78	1
46	1	1	1	50	81	1
47	0	0	0	56	82	1
48	0	0	1	63	82	1
49	1	1	1	65	84	1
50	1	0	1	64	89	1
51	0	1	0	59	99	1

续表

No.	X_RAY	GRADE	STAGE	AGE	ACID	NODES
52	1	1	1	68	126	1
53	1	0	0	61	136	1

上述问题的 SPSS 操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开曲线估计对话框和选入变量。

依次单击菜单“分析→回归分析→二元 logistic 回归”命令，打开“Logistic 回归”对话框。设置“因变量”为 NODES，“协变量”为 X_RAY、GRADE 和 STAGE，连续型自变量为 AGE 和 ACID。

在“Logistic：保存”对话框中选择“概率”、“组成员”和“标准化”。

在“Logistic：选项”对话框中选择“分类图”、“Hosmer-Lemeshow 拟合度”、“个案的列表”和“exp 的 CI 95%”。

03 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
LOGISTIC REGRESSION  NODES
    /METHOD = ENTER X_RAY GRADE STAGE AGE ACID
    /SAVE = PRED PGROUP ZRESID
    /CLASSPLOT /CASEWISE OUTLIER(2)
    /PRINT = GOODFIT CI(95)
    /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

设置完毕后，单击“确定”按钮，执行 logistic 估计操作。

其结果分析如下。

- 基本信息

表 9.45　输出的基本信息

Case Processing Summary

Unweighted Cases[a]	N	Percent
Selected Cases　Included in Analysis	53	100.0
Missing Cases	0	.0
Total	53	100.0
Unselected Cases	0	.0
Total	53	100.0

a．If weight is in effect, see classification table for the total number of cases.

注释：在表 9.45 中，Unweighted Cases 为未加权个案类型，N 为个案数，Percent 为各种类型个案数所占百分比。Included in Analysis 为纳入分析的观察个案数，Missing Cases 为缺失的观察个案数，Unselected Cases 为未纳入分析的观察个案数。

表 9.45 给出纳入分析的观察个体数、缺失的观察个案数和未纳入分析的观察个案数等

基本信息。

- 编码信息

表 9.46　编码信息

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

表 9.46 给出了因变量的原数据编码（Original Value），以及计算分析时编码的信息（Internal Value）。

- 分类表

表 9.47　分类表

Classification Table[a,b]

Observed		Predicted		
		NODES		Percentage Correct
		0	1	
Step 0 NODES	0	33	0S	100.0
	1	20	0	0
Overall Percentage				62.3

a．Constant is included in the model.

b．The cut value is .500

注释：表 9.47 中，Observed 为观察量，Predicted 为预测量（NODES 表示节点数，Percentage Correct 表示修正后的百分比值）。

表 9.47 给出了模型中只有常数项而无自变量时，正确预测百分率为 62.3%。也就是说，原数据的 53 个观察个体中，无淋巴结转移者（NODES=0）有 33 人，有淋巴结转移者（NODES=1）有 20 人，如果每一个体均分类到无淋巴结转移者（NODES=0），则可以得到正确的百分率为 62.3%。

- 模型中只有常数项而无自变量时的回归参数及其检验结果

表 9.48　模型中只有常数项而无自变量时的回归参数及其检验结果

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0　Constant	-.501	.283	3.123	1	.077	.606

表 9.48 给出了模型中只有常数项而无自变量时的回归参数及其检验结果。这里的 B 为 B 检验统计量的值，实际上是 $B=\log it(\hat{p})=\ln\frac{20/53}{1-20/53}=-0.500775\approx-0.501$，S.E 为参

数的渐近标准误差，由 Newton-Raphson 迭代产生的信息矩阵之逆阵的对角元素开方获得。Wald 为卡方检验统计量，为 $=(0.500775/0.283378)^2=0.123$，Sig.=0.077 为 Wald 卡方值 3.123 在自由度为 1 时对应的检验 ρ 值。Exp(B)为 B 检验统计量的指数值。

- 单因变量分析结果

表 9.49　单因变量分析结果

Variables not in the Equation

			Score	df	Sig.
Slep	Variables	X_RAY	11.283	1	.001
0		GRADE	4.075	1	.044
		STAGE	7.438	1	.006
		AGE	1.094	1	.296
		ACID	3.117	1	.077
	Overall Statistics		19.451	5	.002

注释：在表 9.49 中，Score 为因子得分，df 为自由度，Sig.为自由度的熵。X_RAY、GRADE、STAGE、AGE 和 ACID 为自变量。

表 9.49 为单因变量分析结果。在将每个变量放入模型之前，采用得分检验方法，检验某一自变量与因变量之间有无联系。由该结果可初步认为在 0.05 的显著性水平下，变量 X_RAY、GRADE 和 STAGE 与因变量之间的联系有统计学意义；AGE、ACID 与因变量之间的联系无统计学意义。

表 9.49 还给出了 X_RAY、GRADE、STAGE、AGE 和 ACID 5 个自变量全部放入模型后的得分检验结果，得到 $\text{Score}x^2=19.451$，自由度为 $df=0.002$，说明模型全局性检验有统计学意义。

- 模型系数全局性检验结果

表 9.50　模型系数的全局性检验结果

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	22.126	5	.000
	Block	22.126	5	.000
	Model	22.126	5	.000

表 9.50 为模型系数的全局性检验结果。对于 Enter 方法，这三种检验的结果相同，说明至少有一个自变量具有统计学意义。

- 模型摘要

表 9.51　模型摘要

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	48.126[a]	.341	.465

a．Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

表 9.51 为模型摘要，给出了 Cox and Snell 决定系数和 Nagelkerke 决定系数分别为 0.341 和 0.465。

- Hosmer and Lemeshow 检验结果

表 9.52　Hosmer and Lemeshow 检验结果

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.954	8	.652

表 9.52 为 Hosmer and Lemeshow 检验结果，Hosmer and Lemeshow 拟合优度得到检验 ρ 值为 0.652，表明由预测概率获得的期望频数与观察频数之间差异无统计学意义，即模型拟合较好。

- Variables in the Equation 信息

表 9.53　Variables in the Equation 信息

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
							Lower	Upper
Step　X_RAY	2.045	.807	6.421	1	.011	7.732	1.589	37.615
	.761	.771	.976	1	.323	2.141	.473	9.700
1[a]　DRADE	1.564	.774	4.084	1	.043	4.778	1.048	21.783
STAGE	-.069	.058	1.432	1	.231	.933	.833	1.045
AGE	.024	.013	3.423	1	.064	1.025	.999	1.051
ACID	.062	3.460	.000	1	.986	1.064		
Constant								

a．Variable(s) entered on step 1: X_RAY, GRADE, STAGE, AGE, ACID.

注释：表 9.53 中 B 为 B 检验统计量的值，S.E.为参数的渐近标准误差，Wald 为卡方检验统计量，df 为自由度，Sig.为 Wald 卡方值在相应自由度时对应的检验 ρ 值，Exp(B)为 B 检验统计量的指数值，95.0% C.I for EXP（B）为 EXP（B）在 95.0%置信度下的的置信区间（Lower 为置信区间下限，Upper 为置信区间上限）。X_RAY、GRADE、STAGE、AGE 和 ACID 为自变量。

表 9.53 蕴涵着丰富信息。

可以检查所有变量对回归模型的贡献有无统计学意义，由每一个自变量对应的 ρ 值（sig.）可见，在 0.05 显著性水平下，变量 X_RAY 和 STAGE 有统计学意义，ACID 在显著

性水平附近，而变量 GRADE 和 AGE 无统计学意义。

由每个自变量对应的 $\exp(\beta)$，可获得每个自变量对应的优势比 *OR* 值及其 95%的置信区间。

- 观察分组与预测概率分类图

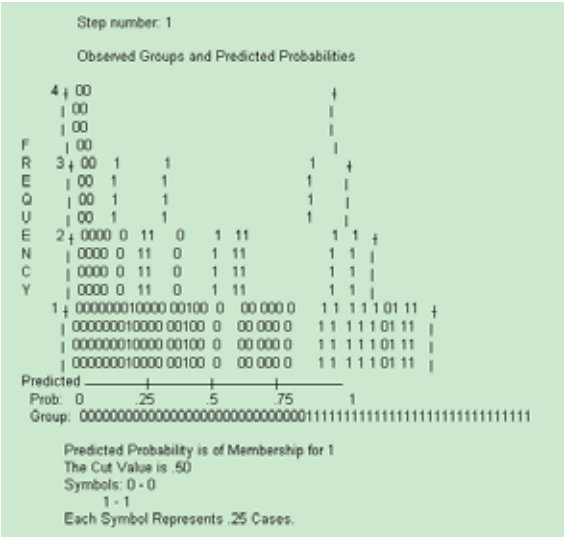


图 9.28　观察分组与预测概率分类图

由图 9.28 可见，在 53 个观察个体（33 个“ 0 ”个体，20 个“ 1 ”个体；结果中每 4 个数字代表 1 个个体）中，大多数“ 0 ”个体在预测概率 0.5 的左边，“ 1 ”个体在预测概率 0.5 的右边，这是分类正确的情况；但预测概率 0.5 左边也有 7 个“ 1 ”个体，右边也有 5 个“ 0 ”个体，这是分类错误的情况。分类基本上呈 U 型，左右数字较多，而中间数字较少。

9.5　本章小结

回归分析是一种非常实用、处理两个及两个以上变量间线性依存关系的统计方法。它的应用范围很广，是统计分析中常用方法之一。回归分析的内容十分丰富，本章介绍了 4 种常用的回归分析方法：线性回归、非线性回归、曲线估计和 logistic 回归。读者在理解回归原理的基础上，应重点掌握 SPSS 回归分析的基本操作和输出结果的相关解释。

第10章 聚类分析

聚类分析又称群分析，是分类学的一种基本方法。通俗地讲，“类”就是由具有相似性的元素构成的集合。聚类分析也是多元统计学中应用极为广泛的一类重要方法。

分类问题在经济、医学以及科学研究中十分常见。例如，超市商品的种类繁多，需要根据商品的用途、价格档次以及产地等多方面的变量因素分成不同的组别，仅利用单变量因素分布不足以全面、综合地描述商品的类别。多种因素同时考虑，对不同品种之间的关系给出定量化的描述，然后指定分组规则，按照实际的需要分类，这时可以利用聚类分析的方法。

从上面的例子可知，聚类分析就是对研究样品或指标进行分类的一种多元统计方法。目前这种方法正处于发展阶段，理论上还不是很完善；但由于它能够解决许多实际问题，因此在很多具体问题以及应用建模中得到了人们的重视，特别是和其他统计方法结合起来使用效果更好。

在应用中，聚类分析还能提供对数据的解释方法，例如统计图形、机理分析等。所以聚类分析过程是一种分类的科学方法，更是一种分类的艺术。

在 SPSS 中进行聚类分析，可以利用“分析→分类”命令实现，如图 10.1 所示。

本章学习要点：

- 聚类分析概述；
- 系统聚类分析；
- 二阶聚类分析；
- K-均值聚类法。

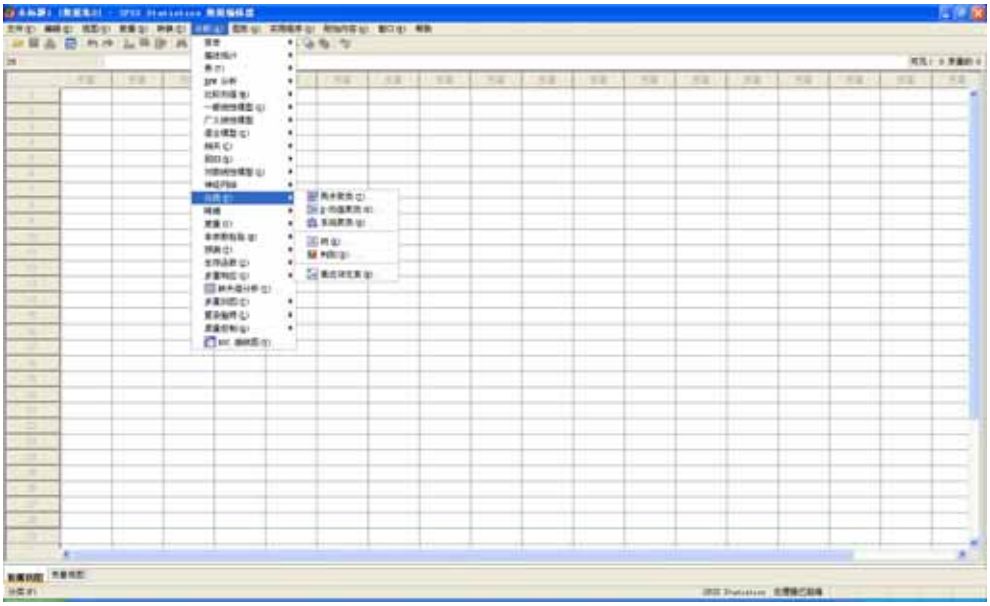


图 10.1　SPSS 聚类分析模块

10.1　聚类分析概述

聚类分析是从事物数值的特征出发对事物进行分类，是数值分类学和多元统计技术相结合的结果，但目前它还是一种较粗糙的、理论并非完善的分析方法。可是，它使用简单，分类效果较好，其内容也在不断丰富中，是一种常用的数据探索性分析工具。

10.1.1　聚类分析的基本原理

聚类分析（Cluster Analysis），又称集群分析，其分析的基本思想是依照事物的数值特征来观察各样品间的亲疏关系。而样品之间的亲疏关系则由样品之间的距离来衡量，一旦样品之间的距离定义之后，就把距离近的样品归为一类。

传统的聚类分析要求聚类变量为数值变量。设 x_{ik} 为第 i 个样品的第 k 个指标，每个样品测量了 q 个变量，则样品 x_i 和 x_j 之间的距离（D_{ij}）的定义为：

$$D_{ij}(q)=\left(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^{q}\right)^{1/q} \tag{10.1}$$

式（10.1）称为明可夫斯基距离（Minkowski distance），其中 n 为大于 0 的正数。

当 q=1 时，$D_{ij}(1)=\left(\sum_{k=1}^{n}|x_{ik}-x_{jk}|\right)$，称为绝对值距离或 Manhattan 距离，在 SPSS 中称为“ block ”。

当 q=2 时，$D_{ij}(2)=(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^{2})^{1/2}$，称为欧几里德距离（Euclidean distance）。

当 $q=\infty$ 时，$D_{ij}(\infty)=(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^{\infty})^{1/\infty}$，称为契比雪夫距离（Chebychev distance）。

也可以用夹角余弦法和相关系数法两种方法来定义变量之间的距离。若变量 x_i 和 x_j 的夹角余弦 C_{ij} 为：

$$C_{ij}=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{[(\sum_{k=1}^{n}x_{ki}{}^{2})(\sum_{k=1}^{n}x_{kj}{}^{2})]^{1/2}} \tag{10.2}$$

变量 x_i 和 x_j 的相关系数 r_{ij} 为：

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)}{[(\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)^{2})(\sum_{k=1}^{n}(x_{kj}-\overline{x}_j)^{2})]^{1/2}} \tag{10.3}$$

C_{ij} 或 r_{ij} 为变量间的相似系数。

则变量间的距离 D_{ij} 由下式定义：

$$D_{ij}=\sqrt{1-C_{ij}^{2}} \tag{10.4}$$

或

$$D_{ij}=\sqrt{1-r_{ij}^{2}} \tag{10.5}$$

聚类分析既可以对样品聚类，又可以对变量聚类；样品聚类也称为 Q 型聚类，变量聚类也称 R 型聚类。根据样本量的大小，可以使用层次聚类（Hierarchical Cluster）或 K-均值聚类（K-Means Cluster）的方法，后者属于一种快速聚类方法。当样本量较大，数值变量和分类变量并存时，也可以使用二步聚类（Two-step Cluster）法。

10.1.2 聚类分析的基本方法

聚类分析是通过将一批数据的个案或者变量的诸多特征，按照关系的远近程度进行分类。如果关系远近程度的定量描述不一样，那么所利用的聚类方法也不一样，可以产生有差别的聚类结果。

本节利用 5 名学生的能力调查数据，简述聚类分析的基本方法，研究聚类分析的基本概念和分类方法。

表 10.1 描述对 08 级本科生中 5 名不同学科的学生能力测试的结果。

根据 10.1 表，可考虑如下几个变量因素：

- 学科

定序型变量，1 代表理科学生，2 代表工科学生，3 代表文科学生。

表 10.1　5 名学生能力测试结果

个　案	学　科	抽象思维	形象思维	创新能力
case1	1	88.00	78.00	3
case2	1	94.00	63.00	2
case3	3	64.00	89.00	1
case4	3	58.00	91.00	1
case5	2	83.00	88.00	2

- 抽象思维

定距型变量，表示进行抽象思维能力测试的得分。

- 学科

定距型变量，表示进行形象思维能力测试的得分。

- 学科

定序型变量，表示创新能力的综合得分，1 表示一般，2 表示好，3 表示很好。

若将 5 名学生分成 3 类，比较明显的是 case1 和 case2 属于第一类，case3 和 case4 属于第二类，case5 单独为第三类，每一子类中的个体特征显然要比另外子类的特征具有更多的相似性。例如 case1 和 case2 的学科相同，抽象思维能力的测试得分都偏高，形象思维能力测试分数中等或偏低，case5 没有分到第一类或者第二类，这是因为在分 3 类的情况下。如果将 5 名学生分成 2 类，case1 和 case2 仍在同一类，case3 和 case4 也在另外的一类，但对于 case5 来讲，似乎不是很好分类。考虑变量特征，从抽象能力测试得分来看，case5 应该和第一类接近，但没有第一类的高；从形象思维能力测试得分来看，case5 和第二类接近，但也没有第二类的高，学科与两类均不同。

从这个问题可以看出，要进行有效的分类，首先应该对个案和个案之间关系的远近程度给予定量描述，同时，还应该考虑聚类方法及其类型。

- 距离描述

在 SPSS 中，我们利用两者之间的距离描述的远近程度来进行具体细分，还可以考虑类与类之间的距离以及个案与类之间的距离等，即考虑如下距离概念及其计算公式：定距型变量的距离、定类型变量的距离、变量和类之间的距离。

- 聚类类型

聚类类型根据分类对象的不同可分为以下两种：

（1）个案聚类

个案聚类也称为 Q 型聚类，是根据被观测个案的各种特征的各个变量值进行分类。例如，上例中考虑的对学生分类就是个案聚类。

（2）变量聚类

变量聚类在统计学中又称为 *R* 型聚类。描述同一问题的特征变量有很多，但在实际问题中，根据我们对问题的关注点不同，在同一问题中不可能将所有的变量进行考虑，应该找出具有代表性同时又是独立的变量进行研究。例如，在回归分析中，变量之间如果产生

共线性可以导致偏回归系数不能真实反映自变量对因变量的影响等。变量分类就是将具有共同特征的变量分为一类。

- 聚类方法

根据聚类分析过程的不同，聚类分析的方法又可以分为以下三种：快速聚类、分层聚类和二步聚类。

10.1.3 不相似测度的度量方法

在聚类分析中，描述个案（或指标）之间或类与个案以及类与类之间关系远近程度是体现两者之间差异程度的一个定量化的描述，聚类分析定义的距离的概念就是描述两者之间差异程度的量，由于这里的距离描述了个案之间差异的量，所以聚类分析中的距离也称为不相似测度或者差异测度。

在聚类分析中定义不相似测度的概念和公式很多，基本方法是将个案看成多维空间的一个点，引用多维空间中点的距离概念定义不相似测度。距离越小，表示两者之间的关系越紧密，相似特征越多，就越有可能分在同一类；相反，距离越大，表示两者之间的差异性越大。下面给出几种主要的定义。

- 定距型变量的不相似测度

如果考虑具有 m 个变量的个案，其变量值都是定距型变量（连续区间），给出两个个案，表示成两个 m 维向量，即 $x=(x_1,x_2,...,x_m)$、 $y=(y_1,y_2,...,y_m)$。

以下为主要的个案 x 和个案 y 的不相似测度公式，定义为 x 和 y 的距离，其中 $\bar{x}$ 和 $\bar{y}$ 分别表示个案 x 和个案 y 的变量值均值。

（1）欧几里德距离（Euclidean distance）

个案 x 和个案 y 的欧几里德距离是两个个案的 m 个变量值之差的平方和的平方根，即：

$$\text{Euclid}(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \tag{10.6}$$

如果同时考虑多个个案的距离，我们可以利用变量值的距离矩阵表示两两之间的距离大小。

（2）平方欧几里德距离（Squared Euclidean distance）

个案 x 和个案 y 的平方欧几里德距离是欧几里德距离的平方，即：

$$\text{SEuclid}(x,y)=\sum_{i=1}^{m}(x_i-y_i)^2 \tag{10.7}$$

（3）夹角余弦距离（Cosine）

个案 x 和个案 y 的夹角余弦的定义是：

$$\text{Cosine}(x,y)=\frac{\sum_{i=1}^{m}(x_i-y_i)^2}{\sqrt{\sum_{i=1}^{m}x_i^{\,2}\sum_{j=1}^{n}y_i^{\,2}}} \tag{10.8}$$

（4）Pearson 相关系数（Pearson correlation）

个案 x 和个案 y 的 Pearson 相关系数的定义是：

$$\text{Pearson}(x,y)=\frac{\sum_{i=1}^{m}(x_i-\overline{x})(y_i-\overline{x})}{\sqrt{\sum_{i=1}^{m}(x_i-\overline{x})^2\sum_{i=1}^{m}(y_i-\overline{y})^2}} \tag{10.9}$$

（5）契比雪夫距离（Chebychev distance）

个案 x 和个案 y 的契比雪夫距离是两个个案变量值之间的距离为其任意变量的最大绝对差值，即：

$$\text{Chebychev}(x,y)=\max|x_i-y_i| \tag{10.10}$$

（6）Block 距离（Block distance）

个案 x 和个案 y 的 Block 距离是两个变量值绝对值之差的和，即：

$$\text{Block}(x,y)=\sum_{i=1}^{m}|x_i-y_i| \tag{10.11}$$

（7）明可夫斯基距离（Minkowski distance）

个案 x 和个案 y 的明可夫斯基距离的数学定义为：

$$\text{Minkowski}(x,y)=p\sqrt{\sum_{i=1}^{m}|x_i-y_i|^p} \tag{10.12}$$

其中：p 为任意实数。

（8）自定义距离（Customized distance）

个案 x 和个案 y 的自定义距离的数学定义为：

$$\text{Customized}(x,y)=r\sqrt{\sum_{i=1}^{m}|x_i-y_i|^p} \tag{10.13}$$

其中：p、r 为任意实数。

- 计数变量的不相似测度

如果个案的变量值都是计数非连续变量（通常是定序型变量），例如上面例子中个案的“学科”变量值是计数非连续变量，则 SPSS 给出了如下个案不相似测度的两种形式。

（1）卡方不相似测度（Chi-Square measure）

个案 x 和个案 y 的卡方测度数学定义为：

$$\text{Chisp}(x,y)=\sqrt{\sum_{i=1}^{m}\frac{(x_i-\overline{x})^2}{\overline{x}}+\sum_{i=1}^{m}\frac{(y_i-\overline{y})^2}{\overline{y}}} \tag{10.14}$$

（2）Phi 方不相似测度（Phi-Square measure）

个案 x 和个案 y 的 Phi 方测度，数学定义为：

$$\mathrm{Phi}(x,y)=\sqrt{\frac{\sum_{i=1}^{m}\frac{(x_i-\overline{x})^2}{\overline{x}}+\sum_{i=1}^{m}\frac{(y_i-\overline{y})^2}{\overline{y}}}{n}} \tag{10.15}$$

其中，n 为个案变量值的总数。

● 二值变量的不相似测度

如果个案的变量值都是二值（Binary）变量，通常是指定类型变量的取值，并且变量只取两个值：0 和 1。例如，问卷调查中很多问题只要求被访者回答“是”与“否”，这时可以用“1”表示“是”，“0”表示“否”，当然也可以利用其他值表示。SPSS 提供了大量和二值变量有关的距离定义，虽然形式有所不同，但大多数和两个个案的匹配频数以及不匹配频数有关，下面首先介绍相关概念。

假设个案 x 和个案 y 的变量值只取 0 或 1，两个个案中相应的变量取值一致的情况越多，说明这两个个案的相似程度越高，表 10.2 称为四格表，表示两个个案变量取值一致的频数和不一致的频数。

表 10.2　变量值匹配四格表

		个案 y	
		1	0
个案 x	1	A	b
	0	C	d

在表 10.2 中，a 是两个个案同时为 1 时的频数，d 是两个个案同时为 0 时的频数，c 是个案 x 为 0 个案 y 为 1 时的频数，b 是个案 x 为 1 个案 y 为 0 时的频数。显然 $a+d$ 反映了两个个案相似的程度，$b+c$ 反映了两个个案差异的程度。

下面利用上述的相似和差异频数来定义各种 SPSS 提供的二值变量不相似测度以及相似测度，定义中出现的 a、b、c、d 的含义均为表 10.2 的四格表所示，n 为个案的变量值的总和。

（1）欧几里德距离（Euclidean distance）

二值变量的欧几里德距离，数学定义为：

$$\mathrm{Euclid}(x,y)=\sqrt{(b+c)} \tag{10.16}$$

（2）平方欧几里德距离（Squared Euclidean distance）

二值变量的平方欧几里德距离，数学定义为：

$$\mathrm{SEuclid}(x,y)=b+c \tag{10.17}$$

（3）简单匹配系数（Simple matching）

二值变量的简单匹配系数，数学定义为：

$$\mathrm{SMatching}(x,y)=\frac{b+c}{a+b+c+d} \tag{10.18}$$

（4）Jaccard 相似性指数

二值变量的相似性指数，数学定义为：

$$\text{Jaccard}(x,y)=\frac{b+c}{a+b+c} \tag{10.19}$$

和简单匹配系数相比，分母中没有考虑两个变量对应变量值同时为 0 的频数，这在医学中比较常用。

（5）Hamann 相似性测度

二值变量的相似性测度，数学定义为：

$$\text{Hamann}(x,y)=\frac{a+d-b-c}{n} \tag{10.20}$$

范围为-1~1。

（6）Lane and Williams 不相似性测度

二值变量的不相似性测度，数学定义为：

$$\text{Lane-Walliams}(x,y)=\frac{b+c}{2a+b+c} \tag{10.21}$$

除了上面列出的二值变量的相关测度以外，还有一些不常用的二值变量的测度，如下所示，用户可以通过查阅 SPSS 帮助文档理解其含义，限于篇幅，这里就不一一介绍了。

（7）不对称指数（Size difference）

（8）不相似测度（Pattern difference）

（9）Dsipersion 相似性指数（Dsipersion）

（10）Shape 相似性指数（Shape）

（11）Pearson 相关系数的二值模拟（Phi4-point correlation）

（12）Lambda 相似性指数（Lambda）

（13）Anderberg'D 相似性指数（Anderberg'D）

（14）Dice 相似性测度（Dice）

（15）Kulczynski 1 相似性指数（Kulczynski 1）

（16）Kulczynski 2 相似性指数（Kulczynski 2）

（17）Ochiai 余弦相似性测度（Ochiai）。

虽然二值不相似的形式有很多种，但一般应用比较广泛的是简单匹配系数（Simple matching）和 Jaccard 相似性指数（Jaccard）两种。

10.2　系统聚类分析

聚类的方法有多种，最常用的是系统聚类分析。根据分层聚类过程的不同，又分为凝聚法和分解法两种。凝聚法的原理是将参与聚类的每个个案（或变量）视为一类，根据两类之间的距离或相似性，逐步合并直到合并为一个大类为止。分解法的原理是将所有个案（或变量）都视为属于一大类，然后根据距离和相似性逐层分解，直到参与聚类的每个个

体自成一类为止。但无论哪种方法，其聚类原则都是近似的聚为一类，即距离最近或最相似的聚为一类；因此，实际上以上两种方法是方向相反的两种聚类过程。

系统聚类分析中的技术问题是描述个案和个案、个案和类以及类与类之间的远近关系，前面已经定义了个案与个案之间的距离，下面继续定义个案与类以及类与类之间的距离，这也是系统聚类分析过程中首先要确定的聚类方法。个案与类以及类与类之间的距离主要有组间平均链锁法（Between-groups linkage）、组内平均链锁法（Whthin-groups linkage）、最近相邻法（Nearest neighbor）、最远相邻法（Furthest neighbor）、重心聚类法（Centroid clustering）、中位数聚类法（Median clustering）和离差平方和法（Ward’s method）。

分层聚类的 SPSS 操作步骤如下。

- 选择分层聚类对话框

依次单击菜单“分析→分类→系统聚类”命令，打开“系统聚类分析”对话框，如图 10.2 所示。



图 10.2 “系统聚类分析”对话框

- 选择变量

在左侧变量列表框中选择变量后，单击 按钮，可以将其移动到“变量”列表框中，如图 10.2 所示。

- “方法”选项

在图 10.2 所示的对话框中，单击“方法”按钮，打开“系统聚类分析：方法”对话框，选择系统聚类中距离的类别，如图 10.3 所示。

在图 10.3 所示的对话框中，SPSS 提供了聚类过程中的聚类方法、度量标准、转换值以及转换度量等选项。

单击“聚类方法”的下拉菜单，在弹出的下拉菜单中显示 SPSS 的几种主要的分层聚类方法，包括组间联接（Between-groups linkage）、组内联接（Whthin-groups linkage）、最近邻元素（Nearest neighbor）、最远邻元素（Furthest neighbor）、质心聚类法（Centroid clustering）、中位数聚类法（Median clustering）和 Ward 法（Ward’s method）。



图 10.3 “ 系统聚类分析：方法 ” 对话框

“ 度量标准 ” 选项区域中给出对聚类个案的变量值距离类型的选择项。首先指定变量值的类型，在聚类分析中根据距离定义的特征，分为区间（Interval）、计数（Count）、二分类（Binary）。单击 “ 区间 ” 单选按钮，激活右侧 “ 区间 ” 参数框，单击参数框中的下拉菜单按钮，弹出连续型定距变量的距离选择项，共有 8 个：Euclidean 距离、平方 Euclidean 距离、余弦、Pearson 相关性、Chebychev 距离、块、Minkowski 距离和设定；单击 “ 计数 ” 单选按钮，激活右侧的 “ 计数 ” 参数框，单击参数框中的下拉菜单按钮，弹出计数变量的距离选择项，共有两个距离选项：卡方度量和 Phi 方度量；单击 “ 二分类 ” 单选按钮，激活右侧的二分类参数框，单击参数框中的下拉菜单按钮，弹出二分类的距离选项，共有 26 个距离选项：Euclidean 距离、平方 Euclidean 距离、尺度差分、模式差别、方差、离散、形状、简单匹配、Phi 4 点相关、Lambda、Anderberg’D、Dice、Hamann、Jaccard、Kulczynski、Kulczynski2、Lane and Williarns、Ochiai、Rogers and Tanimoto、Russel and Rao、Sokal and Sneath1、Sokal and Sneath2、Sokal and Sneath3、Sokal and Sneath4、Sokal and Sneath5 和 Yule’s。

“ 转换值 ” 选项组给出变量的标准化方法的选项。单击 “ 标准化 ” 的下拉菜单，弹出的下拉菜单中显示 SPSS 给出的几种主要的变量值标准化方法。选择标准化方法以后，还要指定标准化是针对变量还是针对个案进行的。表 10.3 是 “ 标准化 ” 下拉菜单中显示的标准化方法选项及其含义。

表 10.3　标准化选项及其含义

选　项	含　义
无	默认选项，不进行标准化
Z 得分	Z 得分值，即在聚类分析之前，将所有变量值按照统计量的标准化方法进行标准化

续表

选　项	含　义
全距-1 到 1	变量值标准化到-1~1 的范围。每个变量值和该变量的极差的比值作为标准化值；如果极差是 0（该变量的所有值相等），则变量值保持不变
全距从 0 到 1	变量值标准化到 0~1 的范围。每个变量值和该变量的极差的比值作为标准化值；如果极差是 0，则变量值设置为 0.5
1 的最大量	变量值标准化到最大值为 1。每个大于 0 的变量值和该变量最大值为该变量的标准化值，如果最大值为 0，则求和最小值的绝对值的比值加 1，小于等于 0 的变量值不变
均值为 1	变量值标准化到平均值为 1。变量值和该变量均值最为该变量的标准化值，如果均值为 0，则对变量值都加 1，使其均值为 1
标准化为 1	数值标准化到变量标准差的范围内。变量值和该变量的样本标准差的比值作为标准化值，如果标准差是 0，则变量值保持不变

“ 转换度量 ”选项组用于确定距离计算完成以后，对距离度量进行转换，提供了三种转换规则的复选项，如表 10.4 所示。

表 10.4　转换测量选项及其含义

选　项	含　义
绝对值	距离值取绝对值。当数值符号表示相关方向且只对负相关关系感兴趣时，使用此方法进行变换
更改符号	改变距离符号。把相似性值变为不相似性值或相反，改变表示距离顺序的方向
重新标度到 0-1 全距	距离与最小值之差，再除以极差得到距离的标准化

- “ 保存 ”选项

聚类分析的结果可以通过建立新变量，保存在当前工作数据文件中。在图 10.2 所示的对话框中，单击“ 保存 ”按钮，打开“ 系统聚类分析：保存 ”对话框，如图 10.4 所示。

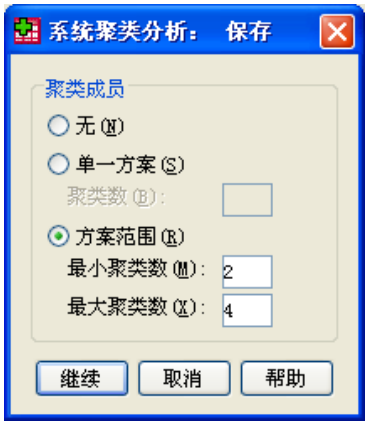


图 10.4 “ 系统聚类分析：保存 ”对话框

在图 10.4 所示的对话框中，“ 聚类成员 ”选项组有三个单选项，其含义如表 10.5 所示。

表 10.5　聚类成员选项及其含义

选　项	含　义
无	默认选项，不建立新变量
单一方案	单一结果，利用新变量表示每个个案聚类后所属的类。在该项的文本框“聚类数”中指定分类数量，可以选择 2~*n* 个整数值，这里 *n* 表示聚类的个案总数
方案范围	选择该项，显示指定聚类数的范围的结果。自动生成若干个新变量，表示聚类为若干个类时，每个个体聚类后所属的类。在该项后面的文本框中指定显示范围，在“最小聚类数”文本框中输入显示聚类的最小类数，在“最大聚类数”文本框中输入显示聚类的最大数

- “统计量”选项

在图 10.2 所示对话框中，单击“统计量”按钮，打开“系统聚类分析：统计量”对话框，如图 10.5 所示。在此对话框中选择输出统计量选项。

在图 10.5 所示对话框中，给出两个复选框和一个选项组。“合并进程表”用于输出凝聚状态表，凝聚状态表显示聚类过程中每一步合并的类或个案，被合并的类或个案之间的距离以及观测量或变量加入到一类的类水平。可以根据凝聚状态表跟踪聚类的合并过程。“相似性矩阵”以矩阵形式显示个案之间的距离或相似性测度值，并在表格中显示距离类型，如果项数很大，输出矩阵的阶数也很大。“聚类成员”选项组中的内容与“系统聚类分析：保存”对话框中的相同，在此不再赘述。

- “绘制”选项

在图 10.2 所示对话框中，单击“绘制”按钮，打开“系统聚类分析：图”对话框，如图 10.6 所示。在该对话框中可以选择输出结果中的图表选项。

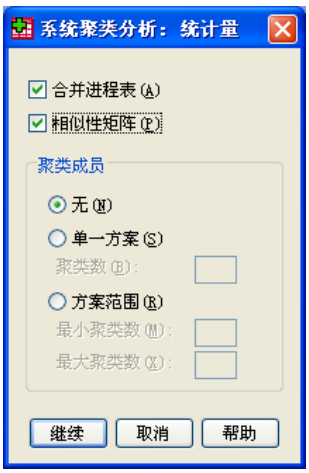


图 10.5　“系统聚类分析：统计量”对话框

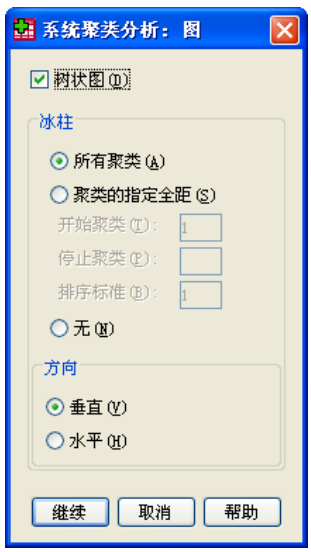


图 10.6　“系统聚类分析：图”对话框

在“系统聚类分析：图”对话框中，可以选择的输出的统计图表有聚类分析树状图和

聚类分析冰柱图。“ 树状图 ” 表示输出结果中显示树形图表。“ 冰柱图 ” 选项组设置关于输出冰柱图的聚类数范围以及图形的形状，冰柱图选项及其含义如表 10.6 所示。

表 10.6　冰柱图选项及其含义

选　项	含　义
无	不生成冰柱图
所有聚类	选择该项，表示输出聚类的每一个阶段的冰柱图。利用冰柱图可以查看聚类分析的整个合并过程。如果个案或者变量数很大，冰柱图形状也会很大，这时可以选择下面的选项，以限制显示的范围
聚类的指定全距	指定显示的聚类数的范围。选择此选项，该选项下的文本框将被激活，要求输入显示聚类的范围

“ 方向 ” 选项组设置作图方向，有两个单选项。如表 10.7 所示。

表 10.7　方向选项及其含义

选　项	含　义
垂直	默认选项，显示纵向冰柱图
水平	选择该项，显示水平冰柱图

【示例 10.1】系统聚类分析。

本例所用数据如表 10.1 中所示。

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开系统聚类分析对话框和选入变量。

依次单击菜单 “ 分析→分类→系统聚类分析 ” 命令，打开 “ 系统聚类分析 ” 对话框。将 “ 学科 ”、“ 形象思维能力测试得分 ”、“ 抽象思维能力测试得分 ” 和 “ 创新能力得分 ” 选入 “ 变量 ” 列表框；将 “ Name ” 选入 “ 标注个案 ” 列表框。

在 “ 分群 ” 选项组选择 “ 个案 ”。

在 “ 输出 ” 选项组选择 “ 统计量 ” 和 “ 绘图 ”。

03 设置。

单击 “ 统计量 ” 按钮，打开 “ 系统聚类分析：统计量 ” 对话框，依次选择 “ 合并进程表 ” 和 “ 相似性矩阵 ”；“ 聚类成员 ” 选择 “ 无 ”。

单击 “ 绘制 ” 按钮，打开 “ 系统聚类分析：图 ” 对话框，选择 “ 树状图 ”；“ 冰柱图 ” 选项组中选择 “ 所有聚类 ”；“ 方向 ” 选项组选择 “ 垂直 ”。

单击 “ 保存 ” 按钮，打开 “ 系统聚类分析：保存 ” 对话框。在 “ 聚类成员 ” 选项组中选择 “ 方案范围 ”，将 “ 最小聚类数 ” 设为 2，“ 最大聚类数 ” 设为 4。

单击 “ 方法 ” 按钮，打开 “ 系统聚类分析：方法 ” 对话框。“ 聚类方法 ” 选择 “ 组间联接 ”；“ 度量标准 ” 选项组中选择 “ 区间 ”，类型为 “ 平方 Euclidean 距离 ”；“ 转换值 ” 选项组 “ 标准化 ” 方法选择 “ Z 得分 ”。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
PROXIMITIES  学科 抽象思维 形象思维 创新能力
    /MATRIX OUT  ('C:\WINDOWS\TEMP\spss2540\spssclus.tmp')
    /VIEW= CASE
    /MEASURE= SEUCLID
    /PRINT  NONE
    /ID= Name
    /STANDARDIZE= VARIABLE Z .
CLUSTER
    /MATRIX IN  ('C:\WINDOWS\TEMP\spss2540\spssclus.tmp')
    /METHOD BAVERAGE
    /ID=Name
    /PRINT SCHEDULE
    /PRINT DISTANCE
    /PLOT DENDROGRAM VICICLE
    /SAVE CLUSTER(2,4) .
ERASE FILE= 'C:\WINDOWS\TEMP\spss2540\spssclus.tmp'.
```

设置完毕后，单击“确定”按钮，执行系统聚类分析操作。

其结果分析如下。

- 距离标准化的比较

表 10.8　距离标准化的比较

Proximity Matrix

Case	Squared Euclidean Distance				
	1:case 1	2:case 2	3:case 3	4:case 4	5:case 5
1:case 1	.000	3.234	12.969	14.651	3.268
2:case 2	3.234	.000	14.102	16.522	6.102
3:case 3	12.969	14.102	.000	.177	3.917
4:case 4	14.651	16.522	.177	.000	5.058
5:case 5	3.268	6.102	3.917	5.058	.000

This is a dissimilarity matrix

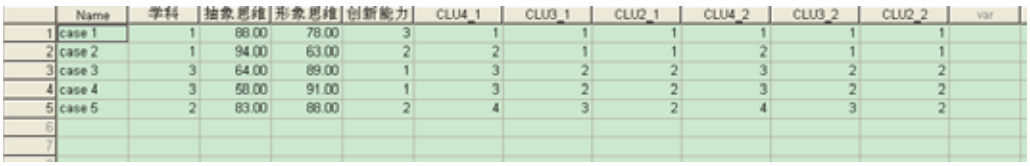
下面结合表 10.1 给出的数据说明该问题。

表 10.8 给出的是欧几里德距离平方和表。考虑 case5 分别和其他 4 个个案的欧几里德距离，要求考虑“抽象思维”和“创新能力”两个变量的变量值，并对数据不进行标准化和利用 Z 得分值进行标准化两种情况做出比较。表 10.8 是计算结果，从结果中可以看出，两种情况的距离顺序不一样，主要是因为“创新能力”数值太小，导致不进行标准化时，“创新能力”的差异几乎没有被考虑到。

- 数据文件输出结果

图 10.7 给出了数据文件中新增聚类变量。CLU4_1 表示分 4 类的情况下，case1、case2、case5 被分在单独的类，而 case3 和 case4 被分在一类；CLU3_1 变量表示分 3 类的情况下，case1 和 case2 在第一类，case3 和 case4 在第二类，case5 单独为一类；CLU2_1 变量表示

分两类的情况下，case1 和 case2 在第一类，case3、case4 和 case5 被分在第二类。



	Name	学科	抽象思维	形象思维	创新能力	CLU4_1	CLU3_1	CLU2_1	CLU4_2	CLU3_2	CLU2_2	var
1	case 1	1	88.00	78.00	3	1	1	1	1	1	1	
2	case 2	1	94.00	63.00	2	2	1	1	2	1	1	
3	case 3	3	64.00	89.00	1	3	2	2	3	2	2	
4	case 4	3	58.00	91.00	1	3	2	2	3	2	2	
5	case 5	2	83.00	88.00	2	4	3	2	4	3	2	
6												
7												

图 10.7　数据文件中新增聚类变量

- 凝聚状态表

表 10.9　凝聚状态表

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	.177	0	0	3
2	1	2	3.234	0	0	4
3	3	5	4.488	1	0	4
4	1	3	11.269	2	3	0

表 10.9 给出本例的凝聚状态表。

- 欧几里德距离矩阵

表 10.10　欧几里德距离矩阵

Proximity Matrix

Case	Euclidean Distance				
	1:case 1	2:case 2	3:case 3	4:case 4	5:case 5
1:case 1	.000	1.798	3.601	3.828	1.808
2:case 2	1.798	.000	3.755	4.065	2.470
3:case 3	3.601	3.755	.000	.421	1.979
4:case 4	3.828	4.065	.421	.000	2.249
5:case 5	1.808	2.470	1.979	2.249	.000

This is a dissimilarity matrix

表 10.10 给出了本例的欧几里德距离矩阵。

- 类成员聚类表

表 10.11 给出了聚类分析的类成员聚类表。由表 10.11 可以看出，当聚成 4 类时，case3 和 case4 聚成一类，其他个案聚成一类；当聚成 3 类时，case1 和 case2 聚成一类，case3 和 case4 聚成一类，case5 自成一类；当聚成两类时，case1 和 case2 聚成一类，其他 3 个个案聚成一类。

表 10.11　聚类分析的类成员聚类表

Cluster Membership

Case	4 Clusters	3 Clusters	2 Clusters
1:case 1	1	1	1
2:case 2	2	1	1
3:case 3	3	2	2
4:case 4	3	2	2
5:case 5	4	3	2

- 树形图

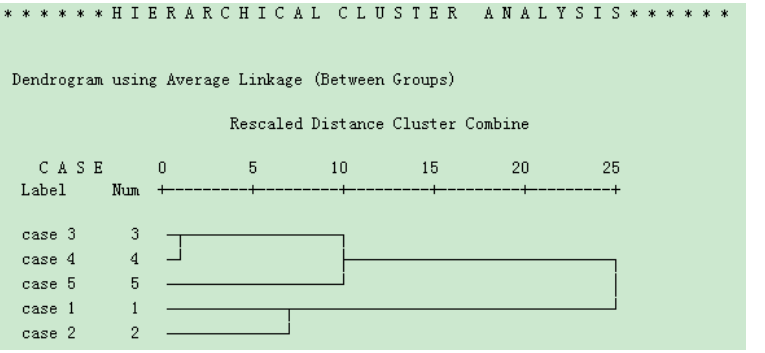


图 10.8　系统聚类分析中的树形图

图 10.8 的聚类分析树形图以水平放置的树形结构呈现了聚类分析中每一次类合并的情况。SPSS 自动将各类间的距离映射到 0~25 之间，并将聚类过程近似地表示在图上。从图中可以看出，case3 和 case4 的距离最近，这一点首先合并成一类；其次 case1 和 case2 合并成一类，这两个个案的距离（1.798）比前两个个案的距离大；再次 case5 与（case3、case4）合并，case5 与（case3、case4）间的距离小于与 case1、case2 间的距离；最后所有个案聚成一类。聚类分析的树形图体现了系统聚类分析方法的主要过程。如果个案数较多或者之间的距离较小，有时就难以分辨聚类的每个过程。

- 冰柱图

表 10.12　系统聚类分析中的冰柱图

Vertical Icicle

Number of clusters	Case								
	5:case 5		4:case 4		3:case 3		2:case 2		1:case 1
1	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X		X	X	X
3	X		X	X	X		X	X	X
4	X		X	X	X		X		X

表 10.12 是以表格的形式呈现出的纵向冰柱图。形状类似于冬天房檐上垂下的冰柱，故以冰柱图命名。观察冰柱图应从最后一行开始。由表 10.12 可知，当聚成 4 类时，case3 和 case4 聚成一类，其他个案自成一类；当聚成 2 类时，case1 和 case2 聚成一类，其他 3 个个案聚成一类。

类成员聚类表、树状图和冰柱图都可以显示不同类数时的个案所属的分类结果，树状图还可以显示聚类的过程，包含更多的信息，但也存在当个案数较多时难以辨别个案与小类之间距离大小的情况，类成员聚类表结构最为简单，3 种表达方式各有优缺点，应用时可根据情况选择显示一种或几种。

10.3 二阶聚类分析

二阶聚类模型是一种新型的系统聚类算法（Herarchical Algorithms），目前一般应用在数据挖掘与多元统计的交叉领域——模式分类中，其算法适合任何尺度的变量。

二阶聚类分析主要是利用距离度量假设聚类模型的变量均为自变量，即假设连续性变量为正态分布，分类变量是多项式。使用经验内部检验方法稳定自变量假设及分布假设的干扰。二阶聚类分析可以产生不同的聚类判别信息、最终聚类的聚类频数、最终聚类的描述性统计量，而且可以产生聚类频数的饼图、条形图和变量的重要性图。

二阶聚类分析具有如下特征：

- 可以处理分类变量和连续性变量。通过假设变量为自变量，综合处理多项正态分布的数据资料；
- 自动选择聚类数。通过比较不同聚类结果的模型选择准则，程序自动确定最优聚类数；
- 有效地分析大样本数据。

二阶聚类分析的计算过程可以分为如下两步：

第一步：构建聚类特征树（Cluster Features Tree）；

第二步：使用聚类法对聚类特征树的节点进行分组。该方法可以产生不同聚类数的指标。通过比较 Schwarz-Bayesian 信息准则或 Akaike 信息准则确定最优类数。

二阶聚类分析的 SPSS 操作步骤如下。

- 选择二阶聚类分析对话框

依次单击菜单“分析→分类→两步聚类”命令，打开“二阶聚类分析”对话框，如图 10.9 所示。

- 确定分类变量和连续型变量

在对话框左侧的变量类表框中选择分类变量，单击 按钮，将其移动到“分类变量”列表框；在左侧的变量列表框中选择聚类的连续型变量，单击 按钮，将其移动到“连续变量”列表框。



图 10.9 “ 二阶聚类分析 ” 对话框

- 选择距离度量

在 “ 距离度量 ” 选项组中确定两个相似性的计算公式，两个选项如表 10.13 所示。

表 10.13　距离度量选项及其含义

选　　项	含　　义
对数相似值	对数似然距离度量，默认选项。对数似然距离度量是基于两聚类之间距离的概率值，假设连续性变量为正态分布，分类变量为多项式分布，并且所有变量以及所有个案之间均假设相互独立。对数似然距离度量既可以用于计算连续型变量，也可以计算分类变量的距离
Euclidean	欧几里德距离度量。两点之间的欧几里德距离定义前面已经介绍过了，两个聚类之间的欧几里德距离是指测度定义为两聚类的类中心的欧几里德距离，类中心定义为类中变量的均值

- 指定聚类数量

在 “ 聚类数量 ” 选项组中指定聚类分析中聚类数的确定方式，对应的两个单选项如表 10.14 所示。

表 10.14　聚类数选项及其含义

选　　项	含　　义
自动确定	系统自动确定聚类数，默认选项。选择该项，“ 最大值 ” 文本框变为可用，在此输入正整数作为聚类分析过程中的系统参考最大聚类数，最大默认值为 15。该方式使用对话框下侧的 “ 聚类准则 ” 选项组指定的聚类准则确定最优的聚类数
指定固定值	固定聚类数。在该单选按钮下的 “ 数量 ” 文本框中输入一个正整数作为聚类数

- 查看连续变量计数

在“连续变量计数”选项组中显示“二阶聚类：选项”对话框中指定的标准化连续型变量的综合信息。参看“二阶聚类：选项”对话框，这里显示的两项内容为：“要标准化的变量”和“假定已标准化的变量”。

- 选择聚类准则

在“聚类准则”选项组中选择确定聚类数的算法，对应的两个单选项为：“施瓦茨贝叶斯准则（BIC，默认选项）”和“Akaike 信息准则（AIC）”。

- “选项”设置

单击图 10.9 所示的图中的“选项”按钮，打开“二阶聚类：选项（Twostep Clusters: Options）”对话框，如图 10.10 所示。

“离群值处理（Outeier Treatment）”选项组用于指定在聚类过程中当产生聚类特征树时奇异值的处理方式。如果聚类特征树在叶的节点处没有获得其他的个案，该树是满的，这时，不存在可以拆分的叶的节点。此选项组中只有一个可选项。



图 10.10 “二阶聚类：选项”对话框

选择“使用噪声处理（Use noise handing）”选项，需要在“百分比”文本框中输入百分比数值。

“内存分配（Memory Allocation）”选项组用于指定聚类计算时的最大内存（MB），指定范围应不小于 4MB。如果计算过程中内存使用超过了指定的最大值，系统将自动将该信息存储至硬盘。注意：当指定内存分配过低的时候，可能会产生聚类的期望类数的计算错误。

在“连续变量的标准化（Standardiyation of Continuous Variables）”选项组中，从左侧的“假定已标准化的变量”列表框中指定需要标准化的连续型变量，将其移动到右侧的“要标准化的变量”列表框。因为聚类算法中要求连续型变量必须是标准化变量，所以应该将

所有连续型变量都移至右侧列表框，以便聚类计算之前标准化所有的连续型变量，这样可以减少计算量，提高聚类效率。

单击“ 高级（Advanced）”按钮，出现高级选项设置，如图 10.11 所示。

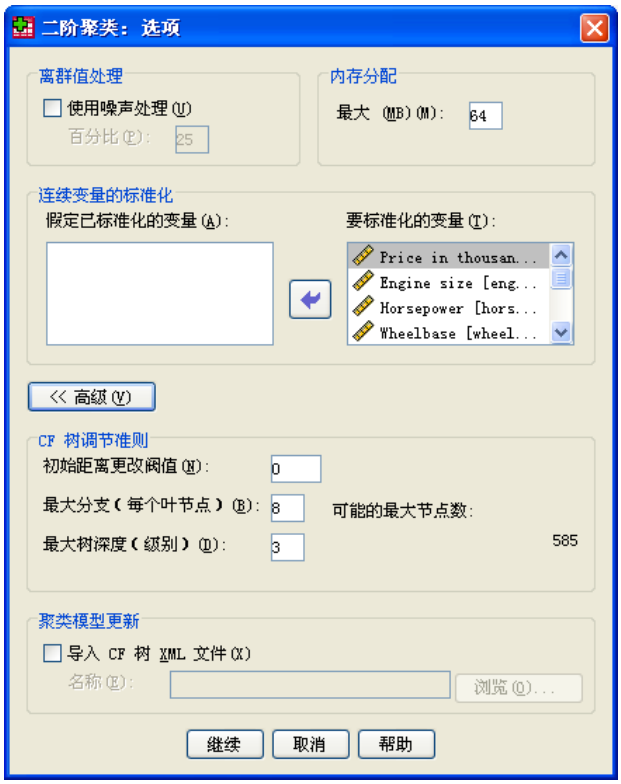


图 10.11 “ 二阶聚类：选项 ”对话框高级设置

在“ 二阶聚类：选项 ”对话框高级设置的“ CF 树调节准则 ”选项组用于指定特征树的聚类计算方法。关于该选项组的选项解释如表 10.15 所示。

表 10.15　二阶聚类选项对话框高级选项及其含义

选　项	含　义
初始距离更改阈值	初始距离改变阈值。如果紧密度小于聚类特征树的初始值，这时可在聚类特征树的叶中插入一个设定的个案而不拆分该叶；如果紧密度大于初始值，则拆分该叶，默认值为 0
最大分支（每个叶节点）	每个叶的子节点的最大分支数，默认值为 8
最大树深度（级别）	最大树深度。聚类特征树的最大级数，默认值为 3
可能的最大节点数	最大可能节点数。聚类过程中可能产生聚类特征树的最大节点数。如果聚类特征树过大，该过程会占用系统过多的内存，且大大增加计算量

“ 聚类模型更新（Cluster Model Update）”选项组用于选择和引入过去分析的模板，同

时更新当前生成的模板。只有一个可选项：导入 CF 树的 XML 文件（Import CFTree XML File）。选择该项，“ 名称 ”文本框被激活，可在此文本框中指定 XML 文件的路径和文件名。

- “ 绘图 ” 设置

单击图 10.9 所示的图中的 “ 绘图（Plots）” 按钮，打开 “ 二阶聚类：绘制（Two Steps Clusten:Plots）” 对话框，如图 10.12 所示。

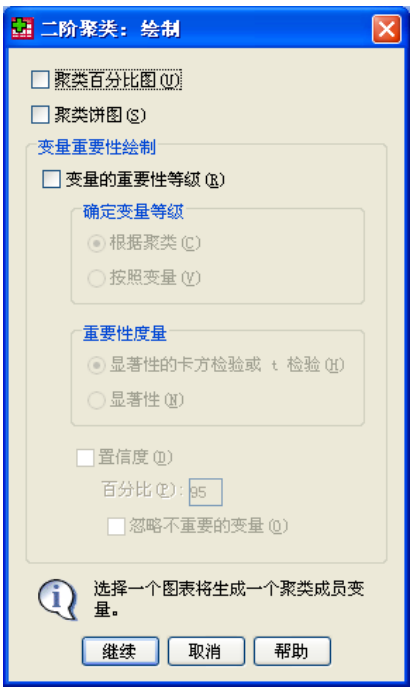


图 10.12 “ 二阶聚类：绘制 ” 对话框

该对话框给出两个选项。如表 10.16 所示。

表 10.16 二阶聚类绘制选项及其含义

选　项	含　义
聚类百分比图	显示每个变量在不同聚类间的差异，对于分类变量，可根据聚类 ID 产生聚类频数的分组条形图，对于每组连续变量，根据聚类的 ID 产生误差条形图
聚类饼图	显示每个聚类观察量总数的饼图和占总观测数百分比数值的饼图

“ 变量重要性绘制（Variable Importance Plot）” 选项组给出了各种不同的图的类型，用于显示各分类中每个变量的重要性，输出结果的顺序是根据每一变量的重要性排序的。

“ 变量的重要性等级（Rank Variable）” 选项组用于指定是根据变量对每一聚类创建图形，还是根据聚类对每一变量创建图形。

“ 重要性度量 ”选项组用于描绘变量重要性测度的计算方法，有两个单选项，如表 10.17 所示。

表 10.17 “ 重要性度量 ” 选项及其含义

选　　项	含　　义
显著性的卡方检验或 T 检验	Pearson 卡方统计量表示聚类变量的重要性，t 值表示连续型变量的重要性
显著性	显著性水平为 1-p，是计算连续型变量的均值和分类变量期望频数的均值相等检验的显著性水平

“ 置信度（Confidence Level）” 复选框设定聚类中每个变量的分布与总体分布的相等性检验的置信度。在 “ 百分比 ” 文本框中输入置信度水平，默认值为 95%，置信度范围设定为大于等于 50 且小于 100 之间。如果图形由变量创建，则置信度水平显示为图形中的垂直线。

- “ 输出 ” 设置

单击图 10.9 中的 “ 输出 ” 按钮，打开 “ 二阶聚类：输出 ” 对话框，如图 10.13 所示。

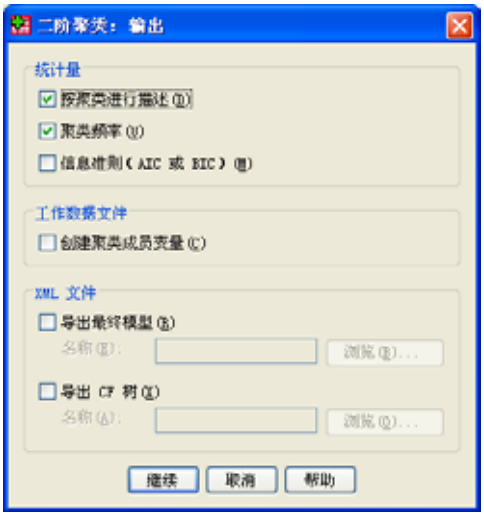


图 10.13 “ 二阶聚类：输出 ” 对话框

在该对话框中 “ 统计量（Statistics）” 选项组给出聚类结果的显示列表选项。选择最终聚类结果的描述性统计量和聚类频数。有 3 个复选框，其含义如表 10.18 所示。

表 10.18　统计量选项及其含义

选　　项	含　　义
按聚类进行描述	选择该项，则在输出结果中显示两个表描述聚类的变量，一个表按聚类列出连续型变量的均值、标准差，另一个表按聚类列出每个分类变量的频数
聚类频率	显示每一分类的观测数目的列表
信息准则	根据主对话框中选择的聚类准则，显示不同聚类中的聚类准则信息 AIC 或 BIC 值的表。该表只适合聚类数由系统自动确定的情况。当聚类数设置为固定值时，该选项将被忽略，不在表中显示

“ 工作数据文件（Working Data File）” 选项组用于显示工作数据文件的聚类变量，有一个复选框：创建聚类成员变量。保存变量到工作数据文件中，新变量中包括每一个个案所属的分类。新创建的变量名为 tsc_n，这里的 n 为一正整数，表示活动数据集的序号。

在“ XML 文件 ”选项组中指定将聚类模型及聚类特征树输出到不同的 XML 格式文件中，有两个复选框，其含义如表 10.19 所示。

表 10.19　XML 文件选项及其含义

选　项	含　义
导出最终模型	聚类模型以 XML 格式输出。选择该项时，在被激活的 “ 名称 ” 文本框中指定保存文件的路径和文件名
导出 CF 树	以外部文件的形式输出聚类特征树。选择该项时，在被激活的 “ 名称 ” 文本框中指定保存文件的路径和文件名

【示例 10.2】不同汽车车型的聚类分析。

现给出 157 个不同汽车的数据。包括汽车生产厂家、汽车型号、各种型号汽车的销售额、价格、燃油效率等相关数据。要求对这 157 个不同型号的汽车个案，利用二阶聚类分析方法进行分类，数据如表 10.20 所示。

表 10.20　157 个不同型号汽车数据信息

manufact	model	sales	resale	type	price	…
Acura	Integra	16.919	16.360	0	21.500	…
Acura	TL	39.384	19.875	0	28.400	
Acura	CL	14.114	18.225	0	#NULL!	
Acura	RL	8.588	29.725	0	42.000	
Audi	A4	20.397	22.255	0	23.990	
…	…	…	…	…	…	

本例 SPSS 操作的具体步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开二阶聚类分析对话框并选入变量。

依次单击菜单“ 分析→分类→两步聚类分析 ”命令，打开“ 二阶聚类分析 ”对话框（如图 10.9 所示）。将“ Vehicle type ”选入“ 分类变量 ”列表框；将从“ Price in thousands through ” 到 “ Fuel efficiency ” 的所有连续型变量移动到 “ 连续变量 ” 列表框。

03 设置。

单击 “ 绘图 ” 按钮，打开 “ 二阶聚类：绘制 ” 对话框（如图 10.13 所示），在该对话框的“ 确定变量等级 ”选项组中选择“ 根据聚类 ”选项，在“ 重要性度量 ”选项组中选择“ 显著性的卡方检验或 T 检验 ” 选项。

单击 “ 输出 ” 按钮，打开 “ 二阶聚类：输出 ” 对话框（如图 10.2 所示）。在 “ 统计量 ”

选项组中选择“按聚类进行描述”和“聚类频率”选项。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
    TWOSTEP CLUSTER
        /CATEGORICAL VARIABLES = type
        /CONTINUOUS VARIABLES = engine_s horsepow wheelbas width length curb_wgt
fuel_cap mpg price
        /DISTANCE LIKELIHOOD
        /NUMCLUSTERS AUTO 15 BIC
        /HANDLENOISE 0
        /MEMALLOCATE 64
        /CRITERIA  INITHRESHOLD (0)  MXBRANCH (8)  MXLEVEL (3)
    /PLOT
        VARCHART COMPARE BYCLUSTER
        CONFIDENCE 95
        /PRINT IC COUNT SUMMARY
        /SAVE VARIABLE=TSC_2372 .
        AIM TSC_2372
        /CATEGORICAL type
        /CONTINUOUS engine_s horsepow wheelbas width length curb_wgt fuel_cap
mpg price
        /PLOT ERRORBAR IMPORTANCE (X=GROUP Y=TEST)
        /CRITERIA ADJUST=BONFERRONI CI=95 SHOWREFLINE=YES HIDENOTSIG=NO .
```

设置完毕后，单击“确定”按钮，执行二阶聚类分析操作。

其结果分析如下。

- 自动聚类表

表 10.21　二阶聚类分析的自动聚类表

Auto-Clustering				
Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[a]	Ratio of Distance Measures[a]
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076

续表

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[a]	Ratio of Distance Measures[a]
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

a．The changes are from the previous number of clusters in the table.

b．The ratios of changes are relative to the change for the two cluster solution

c．The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

表 10.21 所示的是本例的自动聚类表。自动聚类表可表示聚类分析中选择聚类数的过程。第二列的值表示根据 BIC 算法计算出每一种可能的聚类数（共计 15 种可能类数）的 BIC（Schwarz’s Bayesian）。按照将最小的 BIC 值所对应的聚类数作为最优聚类数的确定原则，在本例中，最小的 BIC 值为 885.924，对应的聚类数（第一列）为 3，因此聚类自动分成 3 类。同时，在表的第四列和第五列计算出 BIC 的变化率和距离度量的变化率。一个好的聚类数不仅由 BIC 值决定，还应要求 BIC 的变化率和距离度量的变化率尽可能地相对大一些。综合考虑上述 3 种因素聚类数，聚类数为 3 仍是最优的选择。

- 聚类分布表

表 10.22　二阶聚类分析的聚类分布表

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	62	40.8%	39.5%
	2	39	25.7%	24.8%
	3	51	33.6%	32.5%
	Combined	152	100.0%	96.8%
Excluded Cases		5		3.2%
Total		157		100.0%

表 10.22 所示的是聚类分布表。表中显示每一聚类的频数及其占总个案数和总有效个数的百分比。可以看出，在 157 个个案中，有 152 个参与了聚类过程，62 个分配到第 1 类，39 个分配到第 2 类，51 个分配到第 3 类，有 5 个缺失值。

- 类中心列表

从表 10.23 中可以看出，类型 1 的个案代表价格便宜、小型、燃油效率高的经济型汽车；类型 2 的个案代表价格适中，重型、具有大容量汽缸且燃油效率低的汽车类型；类型 3 代表昂贵、大型和燃油效率适中的汽车类型。

表 10.23　类中心列表

	Price in thousands		Engine siza		Horsepower		Wheeleade		Width		Curb weight		Fuei capacity		Fuel efficiency		Log-transfomed sates	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
聚类	18.80807	6.655822	2.157	.3956	140.71	27.882	102.2891	3.8761	68.490	1.9073	2.82983	.314339	14.831	1.7897	27.47	3.530	3.3550	1.20683
1	26.31998	10.169436	3.520	.9560	186.40	39.730	112.520	9.9483	72.593	4.2335	3.93510	.694206	21.895	4.3670	19.70	3.107	3.9984	.98702
2	37.23648	16.980029	3.659	.9388	213.00	53.827	108.489	6.1224	72.767	2.2653	3.54900	.317543	18.404	2.0235	23.01	2.066	2.7667	1.48915
3	27.33182	14.418669	3.049	1.0498	184.81	56.823	107.414	3.4647	71.089	3.4647	3.37618	.636593	17.959	3.9376	23.84	4.305	3.2890	1.33318
组合																		

注释：表 10.23 所示的是每一聚类中的变量的均值和标准差。Engine size、Horsepower、Wheelbase、Width、Length、Curb weight、Fuel capacity、Fuel efficiency 和 Price in thousands 均为表征汽车性能的参数，分别为发动机大小、马力大小、前后轮轴距、宽度、长度、燃油容量、燃油效率和价格（Mean 为均值，Std.Deviation 为标准差）。

- 分类变量频数表

表 10.24　二阶聚类分析的分类变量频数表

Vehicle type					
		Automobile		Truck	
		Frequency	Percent	Frequency	Percent
Cluster	1	61	54.5%	1	2.5%
	2	0	.0%	39	97.5%
	3	51	45.5%	0	.0%
	Combined	112	100.0%	40	100.0%

表 10.24 所示的是分类变量频数表。分类变量为“ Vehicle type ”，根据分类变量得到的频数进一步说明聚类结果的特征。从表中可以看出，第 2 类全部由客车车型构成，第 3 类全部由轿车车型构成，而第 1 类除了一例外，也全部由轿车构成。

- 变量重要性的聚类条形图

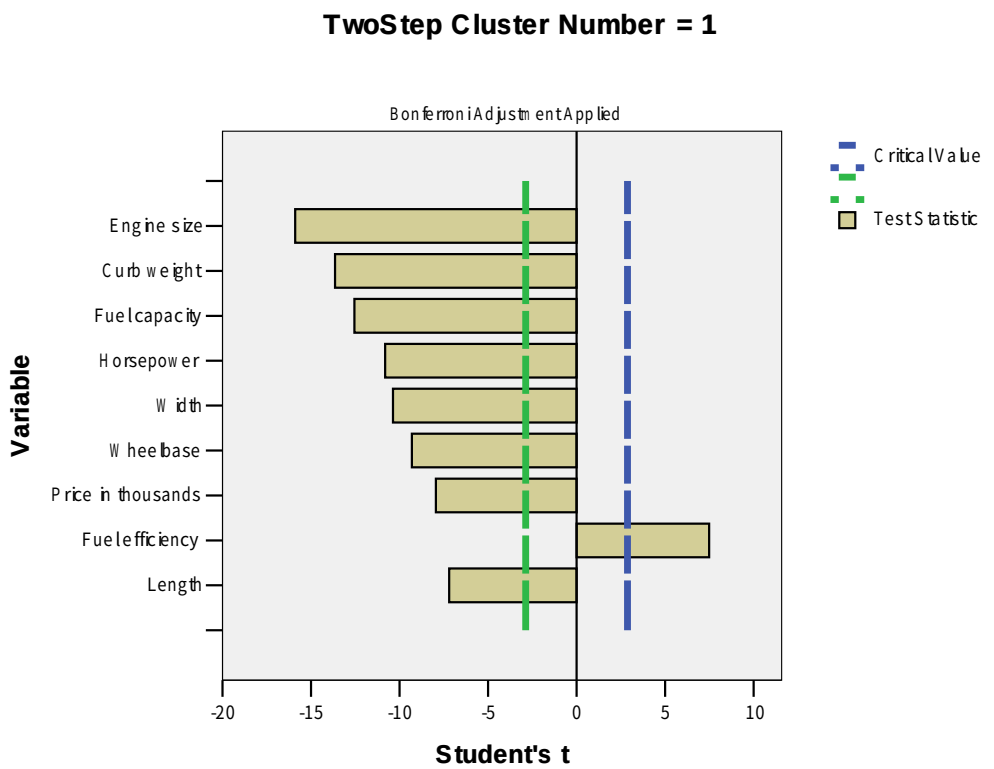


图 10.14　第 1 类的变量重要性的聚类条形图

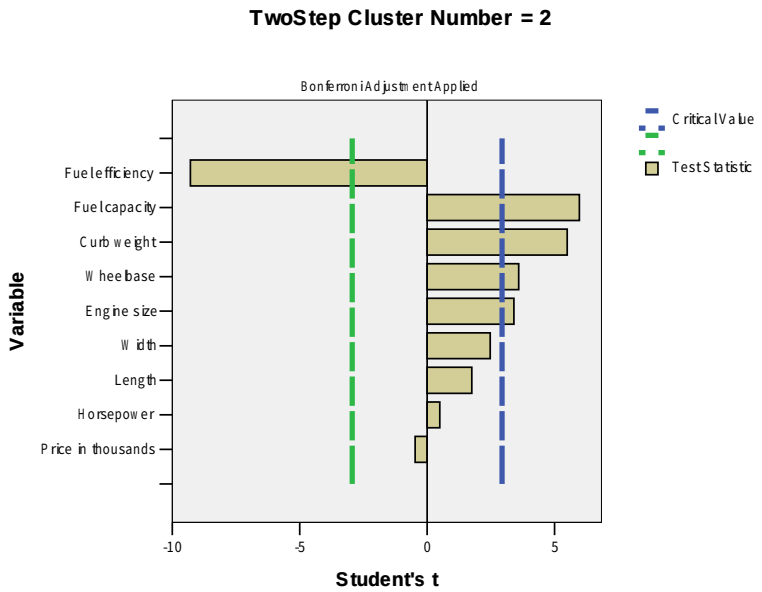


图 10.15　第 2 类的变量重要性的聚类条形图

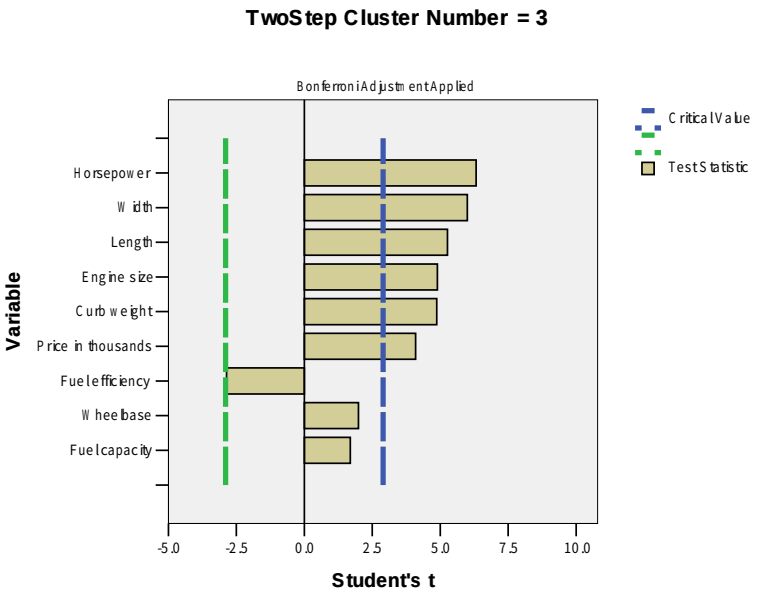


图 10.16　第 3 类的变量重要性的聚类条形图

注释：图 10.14、图 10.15 和图 10.16 分别是变量重要性的 3 个聚类的条形图。Engine size、Horsepower、Wheelbase、Width、Length、Curb weight、Fuel capacity、Fuel efficiency 和 Price in thousands 均为表征汽车性能的参数，分别为发动机大小、马力大小、前后轮轴距、宽度、长度、燃油容量、燃油效率和价格。Critical Value 为每一个变量显著性检验的阈值，Test Statistic 为表示相应变量的 t 统计量。

图 10.13、图 10.14 和图 10.15 分别是变量重要性的 3 个聚类的条形图。变量重要性的聚类条形图是对每一个聚类分别做的变量重要性排序图，变量按照重要性递减的顺序在 *Y* 轴上从上到下排列。图中的两条铅直虚线表示由每一个变量显著性检验的阈值。如果一个变量和其他变量相比有显著差异，它的 *t* 统计量一定位于两条铅直虚线的两侧，图中矩形框的长度和方向表示相应变量的 *t* 统计量，也是该连续型变量的重要性测度。

图 10.13 所示的是第 1 类变量的重要性测度聚类条形图，可以看到该图的所有变量的重要性测度均超过了阈值，故可认为所有连续型变量在形成第 1 类的聚类过程中都是重要的。

负的 *t* 统计量表示该变量的总的均值；反之，正的 *t* 统计量表示该变量在聚类中的均值大于该变量总的均值。例如，对于第一类，燃油效率是唯一一个 *t* 统计量大于 0 的变量，所以，对于第 1 类汽车来说，燃油效率的均值比其他类的要高一些，而该类其他所有变量均值都小于该变量的总平均水平。这些结果和我们在类中心表中观察到的是一致的。

从第 2 类的条形图中可以看到，变量“ Width ”、“ Length ”、“ Horsepower ”、“ Price in thousands ”在形成第 2 类的聚类过程中是不重要的变量。

10.4 K-均值聚类分析

使用系统聚类分析能够得到多个分类解，但也存在执行时间过长的缺点；当给定分类的类数时，使用 K-均值聚类分析过程可以有效地解决该问题，该方法的特点是分类速度快、占用内存少。

使用 K-均值聚类时，可以利用系统默认设置执行命令，也可以对聚类过程设置各种参数，进行人为的干预。例如，可以事先指定样本的分类数，或者指定使聚类过程终止的判据或迭代次数等。

K-均值聚类分析仍是利用距离作为度量个体之间关系紧密程度的指标，并通过指定分类数而求得聚类结果，其基本步骤如下。

01 选择聚类分析的变量和类数。

参与聚类分析的变量必须是数值型变量，为了标识各观测量所属哪一类，还应指定一个表明观测量特征的变量作为标识变量，例如编号、姓名之类的变量；指定的聚类数必须大于等于 2，但不能大于观测量数。

02 确定 *k* 个初始类中心。

如果选择 *n* 个数值型变量进行聚类分析，在给定聚类数为 *k* 的条件下，首先需指定 *k* 个具有代表性的观测量作为聚类的种子。*k* 个事先指定的观测量就是 *k* 个聚类中心点，也称为初始类中心。用户可以选择自己指定或系统指定方式确定类中心点。

03 根据距离最近原则进行分类。

按照距离这 *k* 个类中心距离最近原则，把观测量分配到各类中心所在的类中，形成第一次迭代的 *k* 个分类。

04 根据聚类终止条件进行迭代。

根据组成每一类的观测量计算各变量均值，每一类中的 n 个均值在 n 维空间中又形成了 k 个点，这就是第二次迭代的类中心，按照这种方法依次迭代下去，直到达到指定的迭代次数或终止迭代的判据要求时，迭代停止，聚类结束。

从上述分析过程可以看出，K-均值聚类分析不仅是快速样本聚类过程，而且是一种逐步聚类分析，所谓逐步聚类分析就是先把被聚类对象进行初始分类，然后逐步调整得到最终分类。

SPSS 进行 K-均值聚类分析的操作步骤如下。

● 选择 K-均值聚类分析对话框

依次单击菜单“ 分析→分类→K-均值聚类 ”命令，打开“ K-均值聚类 ”对话框，如图 10.17 所示。

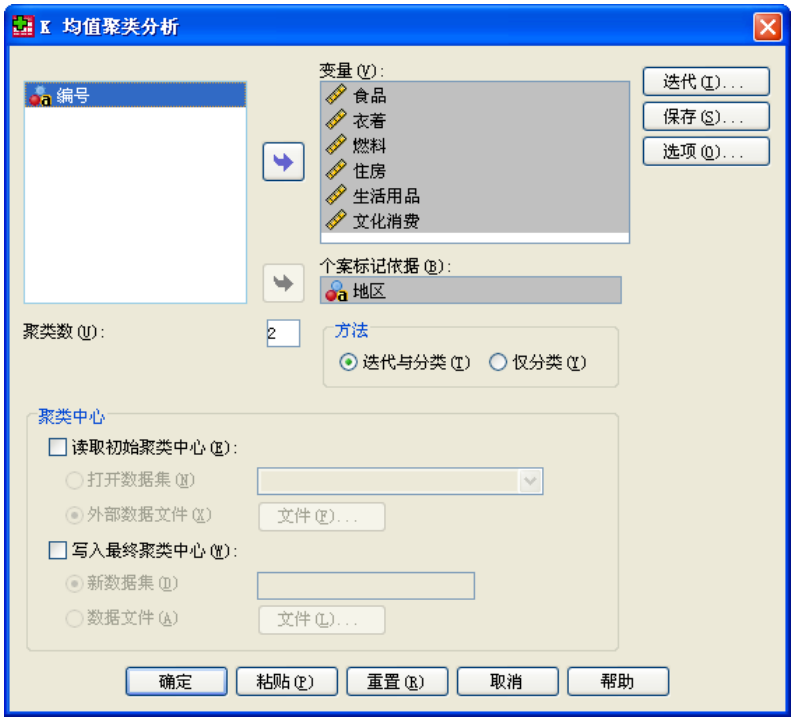


图 10.17 “ K-均值聚类分析 ”对话框

● 选择聚类变量

分别在对话框左侧的变量列表框中选择参与聚类的变量，单击 按钮，将其移动到“ 变量 ”列表框。在对话框左侧的变量列表框选择字符型变量作为标识变量，单击 按钮，将其移动到“ 个案标记依据 ”列表框中。

● 指定聚类过程中类是否调整类中心点

“ 方法 ”选项组中给出聚类过程中两个调整中心点的单选按钮，如表 10.25 所示。

表 10.25　方法选项及其含义

选　项	含　义
迭代与分类	默认选项。表示聚类过程中选择或指定初始类中心点，按照 K-均值的迭代算法不断调整类中心点
仅分类	聚类过程中只使用初始类中心点而不做调整，迭代次数也进行一次

- 聚类中心数据的输入与输出

用户可以指定外部文件或数据集作为初始类中心点，也可以将聚类结果的各类中心数据保存到指定的文件和数据集中。该项功能包含在图 10.16 所示对话框的“聚类中心”选项组中。

如果指定初始中心点，可在图 10.16 所示对话框的“聚类中心”中选项组选择“读取初始聚类中心”；此时，可选择其下的两个单选项，如表 10.26 所示。

表 10.26　读取类中心点选项及其含义

选　项	含　义
打开数据集	如果数据编辑框中打开的数据集多与一个，这时该单选按钮被激活。选择该项后，右侧显示下拉菜单框，打开，显示数据编辑窗口中所有除当前数据集外的其他数据集列表，选择读入初始类中心点的数据集
外部数据文件	选择该项，从外部文件中读入初始类中心点。选中该单选按钮，右侧的“文件”按钮被激活，单击该按钮，在选择文件对话框中选择所需文件后，返回主对话框

如果不选择“读取初始聚类中心”复选框中的选项，则系统自动指定有代表性的初始类中心数据；选择从外部文件或数据集中读入初始类中心点，需要预先建立一个数据文件或数据集，其中个案的数目与要聚类分析的类数相等，每个个案都由参与聚类的变量组成。

如果要将聚类结果中的个类中心数据保存到指定文件，可在图 10.16 所示对话框的“聚类中心”选项组中选择“写入最终聚类中心”复选框，此时，可选如表 10.27 所示的两个单选项。

表 10.27　写入类中心点选项及其含义

选　项	含　义
新数据集	将结果保存在数据编辑窗口中的新建数据集中，在该框右边的文本框中输入数据集名
数据文件	以数据文件的形式保存到指定位置，通过“文件”按钮指定文件保存的路径和文件名

- “迭代”选项

在图 10.16 所示对话框中单击“迭代”按钮，打开“K-均值聚类分析：迭代”对话框，如图 10.17 所示。

该对话框包含两个控制迭代的设置项及类中心的计算选项，如表 10.28 所示。

表 10.28　迭代对话框选项及其含义

选　项	含　义
最大迭代次数	参数文本框。在该文本框中输入 K-均值算法的迭代次数，选择范围为 1~999。默认值为 0。当迭代次数达到最大限定次数时，迭代停止，即使不满足收敛准则也停止
收敛性标准	参数文本框。在该文本框输入 K-均值算法的收敛准则，输入范围在 0~1 之间，且不等于 1。默认值为 0。若输入值为 0.01，表示两次迭代计算的最小的类中心的变化距离在小于初始类中心距离的 1%时迭代停止

如果以上两个文本框都进行了设置，聚类过程中只要满足一个条件迭代就停止；当在“收敛性标准”文本框选择默认值 0 时，只需满足“最大迭代次数”的最大迭代次数为迭代停止的条件。

选择“使用运行均值”复选框，表示在每个个案被分配到一类后立即计算新的类中心；反之，则在完成了所有个案的一次分配后，再计算个类的类中心。注意，对于大数据样本，选择该项会大大增加计算量和计算时间。

- “保存”选项

在图 10.17 所示对话框中单击“保存”按钮，打开“K-均值聚集：保存”对话框，如图 10.19 所示。

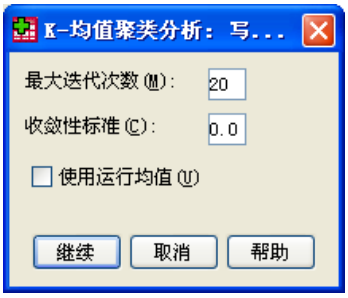


图 10.18　“K-均值聚类：迭代”对话框

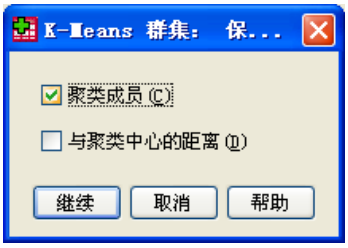


图 10.19　“K-均值聚集：保存”新变量对话框

该对话框只有两个复选框，其选项内容及含义如表 10.29 所示。

表 10.29　保存对话框选项及其含义

选　项	含　义
聚类成员	添加默认变量名为 qcl_1 的新变量，变量值表示聚类结果
与聚类中心的距离	添加默认变量名为 qcl_2 的新变量，变量值表示聚类终止后样本值距所属类中心的欧几里德距离

- “选项”对话框设置

在图 10.17 所示对话框中单击“选项”按钮，打开“K-均值聚类分析：选项”对话框，如图 10.20 所示。

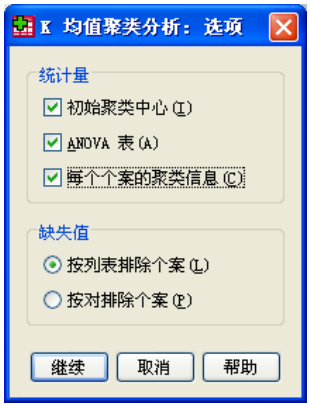


图 10.20 “ K-均值聚类分析：选项 ”对话框

在“ 统计量 ”选项组中选择统计量的输出，有三个可选项目，如表 10.30 所示。

表 10.30 统计量选项及其含义

选　项	含　义
初始聚类中心	默认选项。表示输出初始类中心数据
ANOVA 表	该方差分析是以聚类分析产生的类为控制变量的单因素方差分析，并输出各变量的方差分析表
每个个案的聚类信息	输出每个个案的分类信息，包括分配到哪一类和该个案距所属类中心的距离

在“ 缺失值 ”选项组中包括两个单选项：“ 按列表排除个案 ”和“ 按对排除个案 ”。

【示例 10.3】16 个地区的农民支出情况分析。

表 10.31 是我国 16 个地区的农民在 1982 年支出情况的抽样调查数据的汇总资料，每个地区都调查了反映每人平均生活消费支出情况的 6 个指标。利用 K-均值聚类分析进行分类。

本例 SPSS 操作的具体步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开 K-均值聚类分析对话框并选入变量。

依次单击菜单“ 分析→分类→K-均值聚类分析 ”命令，打开“ K-均值聚类分析 ”对话框（如图 10.16 所示）。将“ 食品 ”、“ 衣着 ”、“ 燃料 ”、“ 住房 ”、“ 生活用品 ”和“ 文化消费 ”选入“ 变量 ”列表框；将“ 地区 ”移动到“ 个案标记依据 ”列表框。

在“ 方法 ”选项组中选择“ 迭代与分类 ”；将“ 聚类数 ”设置为 5，即将 16 个地区分为 5 类。

03 设置。

单击“ 迭代 ”按钮，打开“ K-均值聚类分析：迭代 ”对话框（如图 10.17 所示），在该对话框中，将“ 最大迭代次数 ”设置为 20，“ 收敛性标准 ”设置为 0.01。

表 10.31　16 个地区农民生活水平的调查数据（单位：元）

地　区	食　品	衣　着	燃　料	住　房	生活用品	文化消费
北京	190.33	43.77	9.73	60.54	49.01	9.04
天津	135.20	36.40	10.47	44.16	36.49	3.94
河北	95.21	22.83	9.30	22.44	22.81	2.80
山西	104.78	25.11	6.40	9.89	18.17	3.25
内蒙古	128.41	27.63	8.94	12.58	23.99	3.27
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上海	221.11	38.64	12.53	115.65	50.82	5.89
江苏	144.98	29.12	11.67	42.60	27.30	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5.00
安徽	153.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.26	16.96	19.52	21.75	6.73
江西	140.54	21.50	17.64	19.19	15.97	4.94
山东	115.84	30.26	12.20	33.61	33.77	3.85
河南	101.18	23.26	8.46	20.20	20.50	4.30

单击“ 选项 ”按钮，打开“ K-均值聚类分析：选项 ”对话框（如图 10.19 所示）。在“ 统计量 ” 选项组中选择 “ 初始聚类中心 ”、“ ANOVA 表 ”、“ 每个个案的聚类信息 ” 选项；在 “ 缺失值 ” 选项组选择 “ 按列表排除个案 ”。

在 “ K-均值聚类：保存 ” 对话框（如图 10.18 所示）中，选择 “ 聚类成员 ” 选项。

04 单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
QUICK CLUSTER
    食品 衣着 燃料 住房 生活用品 文化消费
    /MISSING=LISTWISE
    /CRITERIA= CLUSTER(2) MXITER(20) CONVERGE(0.01)
    /METHOD=KMEANS(NOUPDATE)
    /SAVE CLUSTER
    /PRINT ID(地区 ) INITIAL ANOVA CLUSTER DISTAN.
```

设置完毕后，单击 “ 确定 ” 按钮，执行 K-均值聚类分析操作。

其结果分析如下。

- 初始类中心表

本例要求系统自动指定初始类中心点。表 10.32 显示的是系统自动指定的 5 个初始类中心点，初始类中心点分别选择以下地区：北京、天津、河北、上海和江西。可以看出，初始类中心点基本上包括了 16 个地区中高消费地区至低消费地区的各个层次，具有一定的代表性，但一般情况下不是最好，还要通过迭代过程寻找更好的类中心点代替初始类中心点。

表 10.32　K-均值聚类分析的初始类中心数据

Initial Cluster Centers

	Cluster				
	1	2	3	4	5
食品	190.33	135.20	95.21	221.11	140.54
衣着	43.77	36.40	22.83	38.64	21.50
燃料	9.73	10.47	9.30	12.53	17.64
住房	60.54	44.16	22.44	115.65	19.19
生活用品	49.01	36.49	22.81	50.82	15.97
文化消费	9.04	3.94	2.80	5.89	4.94

- 迭代记录

表 10.33　K-均值聚类分析的迭代记录

Iteration History[a]

Iteration	C2hange in Cluster Centers				
	1	2	3	4	5
1	15.478	9.092	11.367	.000	8.437
2	.000	.000	.000	.000	.000

a．Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 36.708.

表 10.33 给出了 5 个类中心每次迭代时的变化。从表中可以看出，第一次迭代后，5 个类的中心点分别变化了 15.487、9.092、11.367、0.000 和 8.437，第一个中心点变化最大；第二次迭代后，5 个类中心点变化均小于指定的收敛准则（0.01），达到聚类结果的要求，聚类分析结束。

- K-均值聚类分析的类成员

表 10.34　K-均值聚类分析的类成员列表

Cluster Membership

Case Number	地　区	Cluster	Distance
1	北京	1	15.478
2	天津	2	9.092
3	河北	3	11.367
4	山西	3	7.825
5	内蒙古	5	19.181
6	辽宁	2	15.678
7	吉林	5	17.880
8	黑龙江	3	13.705
9	上海	4	.000
10	江苏	2	13.564
11	浙江	1	15.478

续表

Case Number	地　　区	Cluster	Distance
12	安徽	5	10.705
13	福建	5	5.146
14	江西	5	8.437
15	山东	2	19.977
16	河南	3	5.278

表 10.34 显示聚类的类成员列表。在类成员列表中，第三列 Cluster 给出了每个个案所属的类，第四列 Distance 显示的是个案和所属类中心的距离。

表中给出了最终 K-均值句类分析的分类结果。从表中可以看出：北京和浙江为第 1 类，天津、辽宁、江苏和山东为第 2 类，河北、山西、黑龙江和河南为第 3 类，上海独自为第 4 类，内蒙古、吉林、安徽、福建和江西为第 5 类。

- K-均值聚类分结束时的类中心数据

表 10.35　K-均值聚类的最后聚类中心

Final Cluster Centers

	Cluster				
	1	2	3	4	5
食品	180.13	135.43	104.35	221.11	145.27
衣着	38.26	32.15	25.19	38.64	25.37
燃料	11.23	13.03	9.35	12.53	15.51
住房	53.83	36.92	16.57	115.65	17.33
生活用品	41.68	34.16	20.81	50.82	21.04
文化消费	7.02	4.25	4.10	5.89	5.31

表 10.35 所示 5 个类的最终聚类中心。可以看出，第 4 类各指数相对来讲较高，第 1 类其次，第 3 类最低。通过最终聚类的类中心数据，可以为后续分析以及解释分类结果，例如分类数的合理确定等问题提供一定的依据。

- 最终聚类的类中心之间的距离

表 10.36　最终聚类的类中心之间的距离

Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		48.877	88.027	74.754	56.221
2	48.877		40.249	117.744	26.572
3	88.027	40.249		156.668	41.409
4	74.754	117.744	156.668		128.418
5	56.221	26.572	41.409	128.418	

表 10.36 所示聚类结束后各个类之间的距离矩阵。从矩阵可以看出，第 3 类和第 4 类之间的距离最大，差距也最大，从类成员列表可知，河北、山西、黑龙江和河南为第 3 类，

上海独自为第 4 类，第 3 类为在国内生活消费水平落后地区，而第 4 类上海，则为全国消费水平最高的地区。

- K-均值聚类分析的方差分析表

表 10.37　K-均值聚类分析的方差分析表

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
食品	3765.241	4	144.552	11	26.048	.000
衣着	106.139	4	20.620	11	5.147	.014
燃料	22.249	4	10.804	11	2.059	.155
住房	2516.642	4	43.493	11	57.863	.000
生活用品	372.115	4	23.410	11	15.896	.000
文化消费	3.735	4	2.268	11	1.647	.232

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

表 10.37 所示各指数在不同类的均值比较情况，各数据项的含义依次为：组间均方、组间自由度、组内均方和组内自由度。通过该方差分析表可以看出 5 类之间的差异。

- K-均值聚类小结

表 10.38　类中含有个案的数量表

Number of Cases in each Cluster

Cluster	1	2.000
	2	4.000
	3	4.000
	4	1.000
	5	5.000
Valid Missing		16.000
		.000

表 10.38 显示 5 个类的类成员情况。第 1 类有 2 个地区，第 2 类有 4 个地区，第 3 类有 4 个地区，第 4 类有 1 个地区，第 5 类有 5 地区，同时还给出了缺失值（Valid Missing）信息。

10.5　本章小结

本章给出了聚类分析的基本原理和几种常用的聚类方法。聚类分析在高级统计分析中已经得到了极为广泛的应用，尤其是在社会学研究领域，聚类分析已经成为了一种不可或缺的分析方法。

聚类分析在本质上是一种分类学方法，虽然从方法上来说并不是很成熟，但几种重要的聚类方法无论从理论上还是在实际应用中都有了很多明显的结果，包括系统聚类分析、二阶聚类分析和 K-均值聚类分析。SPSS 的聚类分析模块给出了以上聚类方法的 3 个过程。读者学习时，要重点掌握和理解聚类分析的基本思想和 SPSS 操作中输出结果所包含的内容。

第11章　判别分析

判别分析（Discriminant Analysis）是类别明确的一种分类技术，它根据观测到的某些指标对所研究的对象进行分类，得到所谓的判别函数，然后再使用判别函数对未知分类的样品进行分类。和聚类分析不同的是，判别分析需要有金标准。

常用的判别分析方法有距离判别、Fisher 的典型判别和 Bayes 判别。距离判别和典型判别对数据分布无严格要求，而 Bayes 判别则要求数据服从多元正态分布。

在 SPSS 中进行判别分析，可以利用“分析→分类→判别分析”命令实现，菜单内容如图 11.1 所示。

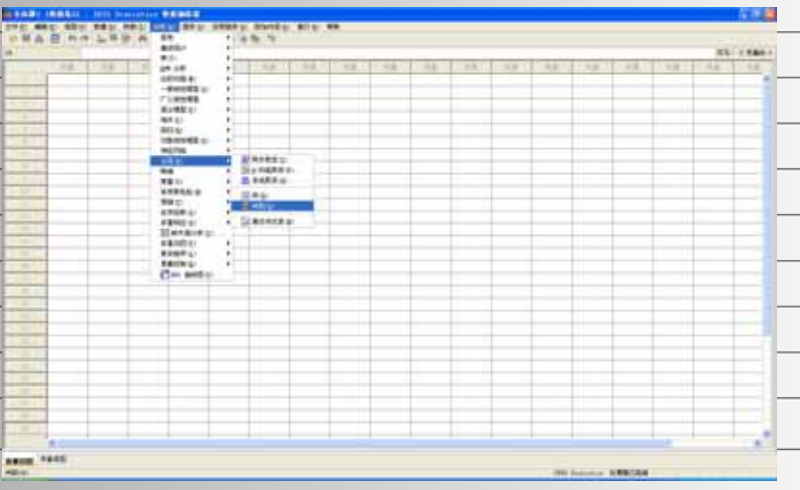


图 11.1　SPSS 判别分析模块

本章学习要点：

- 判别分析的基本原理；
- SPSS 判别分析。

11.1　判别分析的基本原理

判别分析也是一种分类的学科，和聚类分析一样，它已成为高级统计分析不可缺少的一部分。

判别分析是根据观察或测量到的若干变量值判断研究对象如何分类的方法，这一点和第 10 章讨论的聚类分析相似，但聚类分析和判别分析之间又存在一些区别。聚类分析是在未知类别数目的情况下，对样本数据进行分类；而判别分析则是在已知分类数目的情况下，根据一定的指标对不知类别的数据进行分类。

例如，医学诊断中根据化验结果的各个指标判断患者的病症，科学育种时根据种子的外形、营养元素指标、基因特征、成熟期、遗传因素等判断种子的品质。这些问题都可以利用判别分析进行求解。

判别分析的目的是得到体现分类的函数关系式，即判别函数。基本思想是在已知观测对象的分类和特征变量值的前提下，从中筛选出提供较多信息的变量，并建立判别函数；目标是使得到的判别函数在对观测量进行判别其所属类别时的错判率最小。

判别函数的一般形式是：

$$Y = a_1x_1 + a_2x_2 + ...,+a_nx_n \tag{11.1}$$

其中：Y 为判别分数判别值；$x_1, x_2, ..., x_n$ 为反映研究对象特征的变量；$a_1, a_2, ..., a_n$ 为各变量的系数，即判别系数。

11.2　SPSS 判别分析

SPSS 利用“ 判别分析 ”过程进行判别分析操作，下面给出基本操作步骤。

- 选择判别分析对话框

依次单击菜单“ 分析→分类→判别分析 ”命令，打开“ 判别分析 ”对话框，如图 11.2 所示。

- 选择组变量和自变量

在图 11.2 所示对话框左侧变量列表框中选择指定组变量，单击 按钮，将其移动到右侧的“ 分组变量 ”列表框中，并单击“ 定义范围 ”按钮，出现如图 11.3 所示的“ 判别分析：定义范围 ”对话框，在“ 最小值 ”文本框输入该分组变量的最小值，在“ 最大值 ”文本框输入该分组变量的最大值。

在图 11.2 所示对话框左侧变量列表框中选择判别分析的变量，单击 按钮，将其移到右侧的“ 自变量 ”列表框中。



图 11.2 “ 判别分析 ” 对话框

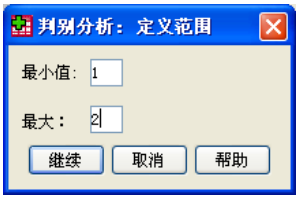


图 11.3 “ 判别分析：定义范围 ” 对话框

● 选择判别分析方法

在图 11.2 所示对话框中，自变量列表框下侧显示两个单选项，用于指定选择判别分析的方法。其内容及含义如表 11.1 所示。

表 11.1　判别分析方法选项及含义

方　法	含　义
一起输入自变量	默认选项。当认为所有自变量都能对观测量特性提供丰富的信息时，使用该选项，选择该项将不加选择地使用所有自变量进行判别分析，建立全模型，且不需要进一步选择
使用步进式方法	当认为不是所有自变量都能对观测量提供丰富的信息时，选择该项，因此需要判别贡献的大小再进行选择。选中该单选按钮，“ 方法 ” 按钮将被激活，可以进一步选择判别分析方法

● 选择变量列表框

如果需要使用一部分个案参与判别函数的推导，而且有一个变量的某个值可以作为这些观测量的标识，则用 “ 选择变量 ” 功能进行选择。方法为在变量列表中选择变量，单击 按钮，将其移动到 “ 选择变量 ” 列表框；然后单击右侧的 “ 值 ” 按钮，显示 “ 判别分析：设置值 ” 对话框，如图 11.4 所示，输入选定变量的值。

分析过程中使用数据文件中的所有合法个案时，此步骤可以省略。

● “ 方法 ” 选项

选择图 11.2 所示的 “ 使用步进式方法 ” 按钮时，“ 方法 ” 按钮被激活，单击 “ 方法 ” 按钮，打开 “ 判别分析：步进法 ” 对话框。如图 11.5 所示。

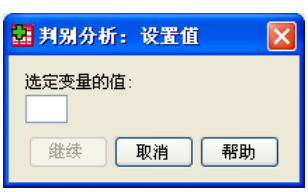


图 11.4 “ 判别分析：设定值 ” 对话框



图 11.5 “ 判别分析：步进法 ” 对话框

在“方法”选项组中进行逐步判别分析的方法，可供选择的判别分析方法有五种，其选项及其含义如表 11.2 所示。

表 11.2　方法选项及其含义

方　法	含　义
Wilks’lambda	默认选项，每步都是 Wilk 的概率量最小的进入判别函数
未解释方差	选择该项，表示每步都是使各类不可解释的方差和最小的变量进入判别函数
Mahalanobis 距离	选择该项，表示每步都使靠得最近的两类间的 Mahalanobis 距离最大的变量进入判别函数
最小 F 值	选择该项，表示每步都使任何两类间的最小的 *F* 值最大的变量进入判别函数
Rao’sV	选择该项，表示每步都使 Rao *V* 统计量产生最大增量的变量进入判别函数，需要在该项下面的“*V* 进入值”文本框中输入这个增量的指定值，当某变量导致的 *V* 值增量大于指定值的变量后进入判别函数

“标准”选项组用于选择逐步判别停止的依据，可供选择的依据如表 11.3 所示。

表 11.3　标准选项及其含义

方　法	含　义
使用 F 值	默认选项。使用 *F* 值是系统默认的依据，当加入一个变量（或剔除一个变量）后，对判别函数中的变量进行方差分析。当计算的 *F* 值大于指定的“进入”值时，该变量保留在函数中，默认“进入”的值是 3.84；当该变量使计算的 *F* 值小于指定的移除值时，该变量从函数中剔除，默认“移除”值为 2.71。即当被加入变量 *F* 的值为 3.84 时，才把该变量加入到模型中，否则变量不能进入模型；或者当要从模型中移除变量 *F* 的值为 2.71 时，该变量才被移出模型，否则模型中的变量不会被移出。设置这两个值时应该注意“进入”值和“移除”值
使用 F 的概率	选择该项，表示用 *F* 检验的概率决定变量是否被加入函数或被剔除，而不是用 *F* 值加入变量的，*F* 值概率的默认值是 0.05，移出变量的 *F* 值概率是 0.10。“移除”值是移出变量的 *F* 值概率，“进入”值是加入变量的 *F* 值概率

“输出”选项组的选项用语显示逐步变量判别法的过程设置。两个复选项及其含义如表 11.4 所示。

表 11.4　输出选项及含义

显示选项	含　义
步进摘要	要求在逐步选择变量过程中的每一步之后显示两个变量的统计量
两两组间距离的 F 值	要求显示两类之间的两两 *F* 值矩阵

- “统计量”选项

在图 11.2 所示对话框单击“统计量”按钮，打开“判别分析：统计量”对话框，如图 11.6 所示。

该对话框给出了输出结果中显示的统计量，在“描述性”选项组中选择对原始数据的描述统计量的输出，其选项及含义如表 11.5 所示。

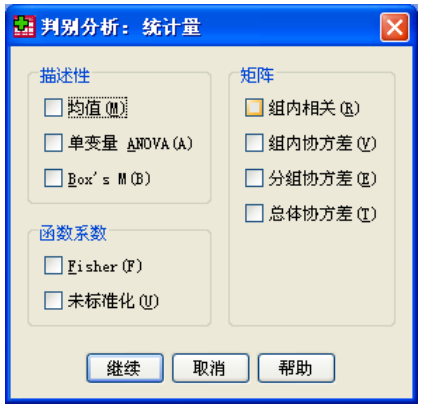


图 11.6 “ 判别分析：统计量 ” 对话框

表 11.5 描述选项及含义

选　项	含　义
均值	选择该项，可以输出各类中各自变量的均值、标准差以及各自变量总样本的均值和标准差
单变量 ANOVA	选择该项，表示对每一类同一自变量均值都相等的假设进行检验，输出单因变量的方差分析结果
Box ’ s M	选择该项，表示对各类的协方差矩阵相等的假设进行检验

“ 函数系数 ” 选项组中给出选择判别函数系数的输出形式的选项，其选项及含义如表 11.6 所示。

表 11.6 函数系数选项及含义

选　项	含　义
Fisher’s	选择该项，表示可以直接用于对新样本进行判别分类的 Fisher 系数，对每一类给出一组系数，并给出该组中判别分数最大的观测量
未标准化	选择该项，表示未经标准化处理的判别系数

在 “ 矩阵 ” 选项组中选择自变量的系数矩阵，有 4 个复选项，如表 11.7 所示。

表 11.7 矩阵选项及含义

选　项	含　义
组内相关	在计算相关矩阵之前将各组协方差矩阵平均后，计算类内相关矩阵
组内协方差	将各组（类）协方差矩阵平均后再计算，区别于总协方差矩阵
分组协方差	协方差矩阵
总体协方差	总样本的协方差矩阵

- “ 分类 ” 选项

在图 11.2 所示对话框单击 “ 分类 ” 按钮，打开 “ 判别分析：分类 ” 对话框，如图 11.7 所示。



图 11.7 “ 判别分析：分类 ” 对话框

该对话框用于指定分类参数和判别结果。下面分别介绍各个选项组的选项。

在 “ 先验概率 ” 选项组中选择先验概率，有两个单选项供选择，如表 11.8 所示。

表 11.8　先验概率选项及含义

选　项	含　义
所有组相等	表示各类先验概率相等，若分为 m 类，则各类先验概率均为 $1/m$
根据组大小计算	表示由各类的样本量计算决定，即各类的先验概率与其样本量成正比

在 “ 使用协方差矩阵 ” 选项组中选择分类使用的协方差矩阵，两个单选项如表 11.9 所示。

表 11.9　使用协方差矩阵选项及含义

选　项	含　义
在组内	选择该项，表示指定使用合并组内协方差矩阵进行分类
分组	选择该项，表示指定使用各组协方差矩阵进行分类，由于分类是根据判别函数而不是根据原始变量，因此该选项不是总等价于二次判别

在 “ 图 ” 选项组中选择要求输出的统计图形，三个单选项如表 11.10 所示。

表 11.10　图选项及含义

选　项	含　义
合并组	选择该项，生成一张包含各类的散点图，该散点图是根据前两个判别函数值做的散点图；如果只有一个判别函数，就输出直方图
分组	选择该项，根据前两个判别函数值对每一类生成一张散点图，分为几类就生成几张散点图；如果只有一个判别函数，就输出直方图
区域图	选择该项，生成用于根据函数值把观测量分到各组中去的边界图，此种统计图把一张图的平面划分出与类数相同的区域，每一类占据一个区，各类的均值在各区中用*号标出；如果仅有一个判别函数则不作此图

在“输出”选项组中选择要求输出的统计图形，三个单选项如表 11.11 所示。

表 11.11　输出选项及含义

选　项	含　义
个案结果	要求输出每个观测量，包括判别分数实际类预测类（根据判别函数求得的分类结果）和后验概率等，选择此项还可以选择其附属选择项，选择“将个案限制在前”选项，并在后面的文本框中输入观测量数 *n*，选择此项时仅对前 *n* 个观测量输出分类结果，观测数量大时可以选择此项
摘要表	要求输出分类的综述表，给出正确分类观测量数（原始类和根据判别函数计算的预测类相同）和错分观测量数及错分率
不考虑该个案时的分类	输出对每个观测量进行分类的结果，所依据的判别是由除该观测量以外的其他测量导出的，也称为交互检验结果

“判别分析：分类”对话框还给出了选择对缺失值的处理方法，即“使用均值替换缺失值”选项，表示用该变量的均值代替缺失值。

- “保存”选项

在图 11.2 所示对话框单击“保存”按钮，打开“判别分析：保存”对话框。如图 11.8 所示。

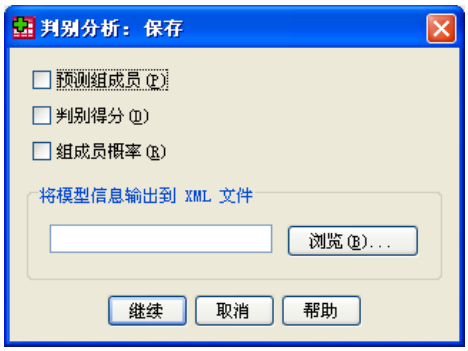


图 11.8　“判别分析：保存”对话框

该对话框用于指定生成并保存在数据文件中的新变量，其中包括表 11.12 所示的三个选项。

表 11.12　保存选项及含义

选　项	含　义
预测组成员	选择该项，要求建立一个新变量预测观测量的分类，是根据判别分数把观测量按后验概率最大指派所属的类，每运行一次判别过程就建立一个，表明使用判别函数预测各观测量属于哪一类的新变量。第 1 次运行建立新变量的变量名为 dis_1，如果在工作数据文件中不把前一次建立的新变量删除，则第 *n* 次运行判别过程建立的新变量默认的变量名为 dis_*n*

续表

选　项	含　义
判别得分	选择该项，要求建立表明判别分数的新变量，该分数是由未标准化的判别系数乘自变量的值，将这些乘积求和后加上常数得来的。每次运行判别过程都给出一组表明判别分数的新变量，建立几个判别函数就有几个判别分数变量参与分析的观测量，共分为 m 类，则建立 m 个点，那么判别函数指定该选择项就可以生成 $m-1$ 个表明判别分数的新变量
组成员概率	选择该项，要求建立新变量表明观测量属于某一类的概率。如果有 m 类，则对一个观测量就会给出 m 个概率值，因此建立 m 个新变量

【示例 11】对一组体检数据的判别分析。

为研究舒张期血压和血浆胆固醇对冠心病的作用，某医生测定了 15 例 50~59 岁冠心病人和 16 例正常人的舒张压和胆固醇指标，数据如表 11.13 所示。

表 11.13　31 例舒张压和胆固醇指标数据

冠心病人组			正常人组		
编号	舒张压	胆固醇	编号	舒张压	胆固醇
1	9.86	5.18	1	10.66	2.07
2	13.33	3.73	2	12.53	4.45
3	14.66	3.89	3	13.33	3.06
4	9.33	7.10	4	9.33	3.94
5	12.80	5.49	5	10.66	4.45
6	10.66	4.09	6	10.66	4.92
7	10.66	4.45	7	9.33	3.68
8	13.33	3.63	8	10.66	2.77
9	13.33	5.96	9	10.66	3.21
10	13.33	5.70	10	10.66	5.02
11	12.00	6.19	11	10.40	3.94
12	14.66	4.01	12	9.33	4.92
13	13.33	4.01	13	10.66	2.69
14	12.80	3.63	14	10.66	2.43
15	13.33	5.96	15	11.20	3.42
			16	9.33	3.63

本例 SPSS 操作的具体步骤如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开判别分析对话框和选入变量。

依次单击菜单“分析→分类→判别分析”命令，打开“判别分析”对话框（如图 11.2 所示）。将变量“组别”选入“分组变量”列表框；单击“定义范围”按钮，在“最小值”文本框中输入 1，在“最大值”文本框中输入 2；将“舒张压”和“胆固醇”移动到“自变量”列表框；选择“一起输入自变量”判别方式作为判别分析的方法。

03 设置。

单击“统计量”按钮，打开“判别分析：统计量”对话框（如图 11.6 所示），在“描述性”选项组中选择“均值”；在“函数系数”选项组中选择“未标准化”。

单击“分类”按钮，打开“判别分析：分类”对话框（如图 11.7 所示）。在“先验概率”选项组中选择“所有组相等”选项；在“输出”选项组中选择“个案结果”和“摘要表”。

单击“保存”按钮，打开“判别分析：保存”对话框（如图 11.8 所示），选择“预测组成员”选项。

其他各选项采用默认值。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
DISCRIMINANT
    /GROUPS=组别(1 2)
    /VARIABLES=舒张压 胆固醇
    /ANALYSIS ALL
    /SAVE=CLASS
    /PRIORS  EQUAL
    /STATISTICS=MEAN STDDEV RAW TABLE
    /PLOT=COMBINED
    /PLOT=CASES
    /CLASSIFY=NONMISSING POOLED .
```

设置完毕后，单击“确定”按钮，执行判别分析操作。

得到其结果分析如下所示。

- 分析个案综合统计量

表 11.14　分析个案综合统计表

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid	Missing or out-of-range group codes	31	100.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
		0	.0
Excluded	Total	31	.100.0

注释：表 11.14 所示为系统处理的数据简明表中的数据。其中 Unweighted Casesa 为未加权个案类型，N 为个案数，Percent 为各种类型个案数所占百分比。Excluded 为剔除的观察个案数，Missing or out-of-range group codes 为有缺失值或超出取值范围的个案数，At least one missing discriminating variable 为至少有一个缺失值的个案数，Both missing or out-of-range groups codes and at least one missing discriminating variable 为有两个缺失值或超出取值范围和至少有一个缺失值的个案数。

表 11.14 所示为系统处理的数据简明表中的数据，按变量“ 组别 ”分组共有 31 个样本作为判别基础数据进入分析，其中第 1 组 15 例，第 2 组 16 例。

- 分组统计表

表 11.15　分组统计表

Group Statistics

组　别		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
1.00	舒张压	12.3794	1.64999	16	16.00
	胆固醇	4.6931	1.29614	16	16.00
2.00	舒张压	10.6267	1.13528	15	15.00
	胆固醇	3.7687	.85021	15	15.00
Total	舒张压	11.5313	1.65996	31	31.00
	胆固醇	4.2458	1.18231	31	31.00

表 11.15 为分组统计表。表中给出分组变量和合计的均值（means）、标准差（standard deviations）和有效个案的例数。

- 典型判别函数的特征值类

表 11.16　典型判别函数的特征值类

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.713[a]	100.0	100.0	.645

a．First 1 canonical discriminant functions were used in the analysis.

表 11.16 是典型判别函数的特征值表，其特征值（Eigenvalues）为组间平方和与组内平方和之比，计算得 0.713，典型相关系数（Canonical Correlation）为 0.645。

- Wilks λ 检验结果

表 11.17　Wilks λ 检验结果

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.584	15.071	2	.001

表 11.17 是典型判别分析的 Wilks λ 检验结果。其中 Wilks λ 值为 0.584，卡方检验统计量的观测值为 15.071，概率 ρ 值为 0.001，小于 0.05。

- 标号化典型判别函数系数

表 11.18　标准化典型判别函数系数

Structure Matrix

	Function
	1
舒张压	.873
胆固醇	.669

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within functio

表 11.18 所示是标准化典型判别函数的系数，其标准化函数为 $ZFunc = 0.873x_1 + 0.669x_2$。

根据判别方程的标准化函数，确定各变量对结果的作用大小，本例舒张压的标准化系数为 0.873，大于胆固醇的标准化系数 0.669，因而舒张压对冠心病的影响作用大于胆固醇。

- 典型判别函数系数

表 11.19　典型判别函数系数

Canonical Discriminant Function Coefficients

	Function
	1
舒张压	.613
胆固醇	.606
(Constant)	-9.638

Unstandardized coefficients

表 11.19 是典型判别函数的系数，其标准化函数为 $ZFunc = -9.638 + 0.613x_1 + 0.606x_2$。

- 分类过程摘要

表 11.20　分类过程摘要

Classification Processing Summary

Processed		31
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		31

注释：表 11.20 中，显示参与分类的个案数和剔除数。其中 Excluded 为剔除的观察个案数，Missing or out-of-range group codes 为有缺失值或超出取值范围的个案数，At least one missing discriminating variable 为至少有一个缺失值的个案数，Used in Output 为在输出时用到的个案数。

表 11.20 为分类过程摘要，显示参与分类的个案例数和剔除的例数。

- 个案统计量表

表 11.21　逐步判别分析过程

Casewise Statistics

Case Number		Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d \| G=g)		P(G=g \| D=d)	Squared Mahalanobis Distance to Centroid	Group	P(G=g \| D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df						
Original	1	1	2*	.699	1	.669	.149	1	.331	1.557	-.457
	2	1	1	.999	1	.791	.000	2	.209	2.665	.789
	3	1	1	.363	1	.944	.828	2	.056	6.473	1.701
Original	4	1	1	.683	1	.661	.167	2	.339	1.503	.383
	5	1	1	.459	1	.927	.549	2	.073	5.642	1.532
	6	1	2*	.830	1	.728	.046	1	.272	2.013	-.628
	7	1	2*	.665	1	.652	.188	1	.348	1.442	-.410
	8	1	1	.950	1	.774	.004	2	.226	2.471	.728
	9	1	1	.177	1	.972	1.824	2	.028	8.909	2.141
	10	1	1	.233	1	.964	1.423	2	.036	7.993	1.984
	11	1	1	.499	1	.920	.456	2	.080	5.335	1.466
	12	1	1	.326	1	.950	.966	2	.050	6.849	1.773
	13	1	1	.867	1	.833	.028	2	.167	3.248	.959
	14	1	1	.699	1	.669	.150	2	.331	1.555	.404
	15	1	1	.177	1	.972	1.824	2	.028	8.909	2.141
	16	1	2*	.313	1	.952	1.020	1	.048	6.991	-1.853
	17	2	1*	.956	1	.776	.003	2	.224	2.494	.736
	18	2	1*	.683	1	.661	.167	2	.339	1.503	.383
	19	2	2	.490	1	.922	.477	1	.078	5.404	-1.534
	20	2	2	.665	1	.652	.188	1	.348	1.442	-.410
	21	2	2	.472	1	.540	.516	1	.460	.838	-.125
	22	2	2	.396	1	.938	.719	1	.062	6.162	-1.692
	23	2	2	.558	1	.908	.342	1	.092	4.926	-1.429
	24	2	2	.750	1	.865	.101	1	.135	3.813	-1.162
	25	2	2	.436	1	.515	.607	1	.485	.731	-.064
	26	2	2	.972	1	.801	.001	1	.199	2.786	-.878
	27	2	2	.923	1	.816	.009	1	.184	2.994	-.940
	28	2	2	.526	1	.915	.402	1	.085	5.144	-1.477
	29	2	2	.429	1	.933	.626	1	.067	5.884	-1.635
	30	2	2	.889	1	.752	.020	1	.248	2.234	-.704
	31	2	2	.380	1	.941	.772	1	.059	6.313	-1.722

**. Misclassified case

表 11.21 为原始数据逐一回归的判别结果和预测分类结果的显示，其中病人组有 3 人被错判（标注*者，编号为 1、6、7），正常人组有 3 人被错判（标注*者，编号为 16、17、18）。

- 分类结果

表 11.22　分类结果

Classification Results[a]

组　别			Predicted Group Membership		Total
			1.00	2.00	
Original	Count	1.00	12	4	16
		2.00	2	13	15
	%	1.00	75.0	25.0	100.0
		2.00	13.3	86.7	100.0

a．80.6% of original grouped cases correctly classified.

表 11.22 给出分类结果。最后系统对判别的情况做评价，即病人组判断正确率为 75.0%，正常人组的正确率判断为 86.7%，总判断正确率为 80.65%，说明该判别分析的正确率较高。

11.3　本章小结

本章内容为判别分析及其 SPSS 操作。判别分析是多元统计分析中判别样本所属类别的一种重要方法，在社会科学和医学卫生领域有非常广泛的应用。读者应了解判别分析和聚类分析之间的区别与联系，掌握 SPSS 判别分析的基本操作和结果解释。

第12章　因子分析与主成分分析

主成分分析（Primary Component Analysis）主要是通过降维过程，将多个相关联的数值指标转化为少数几个互不相关的综合指标的统计方法，即用较少的指标来代替和综合反映原来较多的信息，这些综合后的指标就是原来多指标的主要成分。主成分分析作为一种探索性的技术，是在分析者进行多元数据分析之前用来分析数据，让自己对数据有一个大致的了解，这对于实际应用是非常重要的。因子分析（Factor Analysis）是主成分分析的推广和发展，也是利用降维方法进行统计分析的一种多元统计方法。因子分析研究相关矩阵或协方差的内部依赖关系，由于它将多个变量综合为少数几个因子，以再现原始变量与因子之间的相互关系，故得到了广泛的应用。

在 SPSS 中进行因子分析和主成分分析，可以利用“分析→降维→因子分析”命令实现，菜单内容如图 12.1 所示。



图 12.1　SPSS 因子分析模块

本章学习要点：

- 基本原理；
- 因子分析；
- 主成分分析。

12.1 基本原理

因子分析和主成分分析是不同的两种方法，但分析过程极为相似，且它们在 SPSS 中都是采用 Factor Analysis 过程拟合。下面先分析其基本原理，然后再介绍具有共同性的一些问题。

12.1.1 主成分分析

主成分分析只是一种中间手段，其背景是研究中经常会遇到多指标的问题，这些指标间往往存在一定的相关，直接纳入分析不仅复杂，使变量之间难以取舍，而且可能因多元线性而无法得出正确结论。主成分分析的目的就是运用线性变换，将原来的多个指标组合成相互独立的少数几个能充分反映母体信息的指标，从而在不丢掉主要信息的前提下，避开了变量之间共线性的问题，便于进一步分析。主成分分析不能被看作是研究的结果，而应继续采用其他多元统计方法以解决实际问题。

在主成分分析中，提取出的每个主成分都是原来多个指标的线性组合，比如有两个原始变量 x_1 和 x_2，则一共可提取出如下两个主成分：

$$\begin{cases} PCA_1 = a_{11}x_1 + a_{21}x_2 \\ PCA_2 = a_{12}x_1 + a_{22}x_2 \end{cases}$$

原则上如果有 n 个变量，则最多可提取出 n 个主成分，但如果将它们全部提取出来就失去了该方法简化数据的实际意义。多数情况下提取出前 2~3 个主成分就已包含了 90%以上的信息，其他的可以忽略不计。

提取出的主成分只要能包含主要信息即可，不一定要有准确的实际含义。

所以对于某一问题同时要考虑几个因素时，并不对这些因素单个处理，而是将它们综合起来处理，用数据的本身来寻找对全部方差有影响力的方向、第二有影响力的方向等，当然希望能够找到一两个方向能解释大部分数据中的变异，从而得到“降低原始数据中的维数”的效果，这就是主成分分析的核心概念。

12.1.2 因子分析

“主成分分析”的目的是希望用较少的变量去解释原始数据中的大部分变异，这些变量也就是利用主成分分析法整理而成的整体性指标。而“因子分析”也是希望能够降低变量的数目，但不同的是，只想在一群具有相关性且难以解释的数据中，找出几个在概念上有意义的、并且彼此之间近于独立、可以影响原始数据的共同因素。

因子分析也是一种将多变量化简的技术，它可以被看成是主成分分析的推广。因子分析的目的是分解原始变量，从中归纳出潜在的“类别”，相关性较强的指标归为一类，不

同类间变量的相关性则降低。每一类变量代表了一个“ 共同因子 ”，即一种内在结构、因子分析就是要寻找该结构。

因子分析一般要求提取出的公因子有实际含义，如果分析中各因子难以找到合适的意义，则可以运用适当的旋转，以改变信息量在不同因子上的分析，最终方便对结果的解释。

12.1.3　方法用途

主成分分析与因子分析的用途如下。

- 解决共线性问题：如前所述，利用主成分分析提取出主要信息，然后使用提取出的主成分代替原变量进行分析，就可以避开原变量的共线性问题。
- 评估问卷的结构效度：运用因子分析出的问卷中哪些问题用于研究那些潜在特征（因子），从而得出该问卷结构效度如何的评估。这是社会学和流行病学调查中常用的方法。
- 寻找变量之间的潜在结构：许多变量是无法直接观测到的，它们往往需要用一系列可直接观测的相关变量来间接反映。运用因子分析，就可以将这些变量潜在的结构推导出来加以利用。
- 内在结构证实：有的情况下，研究者根据某些理论或其他知识对可能的内在结构进行了假设，此时可利用因子分析来验证该假设是否成立，这种因子分析也被称为证实性因子分析，在心理学研究中较为常见。

12.1.4　使用条件

主成分分析的目的是提取信息，对样本量没有太严格的要求。因子分析则不可，它更多的是寻找内在结构，因此要求样本量比较充足，否则可能无法得到稳定和准确的结果。按照 Gorsuch（1983 年）的观点，因子分析时的样本量要求如下：

- 样本量与变量数的比例应在 5:10 以上，实际上理想的样本量应为变量数的 1~25 倍，但这很难做到。5~10 倍之间虽略显不足，但一般都能得到较好的结果。总样本量不得少于 100，而且原则上越大越好。
- 以上为对样本量的经验需求，但无论采用的样本有多充分，得到的分析结果都应当在后续工作中继续加以验证。

除了样本量以外，既然要从许多变量中提取共同因素，主成分分析、因子分析有一个默认的前提条件就是各变量间必须有相关性，否则各变量间没有共享信息，就不应当有公因子需要提取，自然也谈不上使用该方法。这是主成分分析和因子分析最为严格的前提要求。具体在该条件的判断上，除了根据专业知识来估计外，还可以使用 KMO 统计量和 Bartlett’s 球形检验加以判定。

- KMO 统计量

用于研究变量之间的偏相关性，它是比较各变量之间的简单相关和偏相关的大小，取值范围在 0~1 之间。如果各变量之间存在内在关系，则由于计算偏相关时控制其他因素就会同时控制潜在变量，导致偏相关系数远远小于简单相关系数，此时 KMO 统计量接近 1，做因子分析的效果最好。一般认为当 KMO 大于 0.90 时效果较佳，0.70 以上时效果尚可，0.60 时效果很差，0.50 及以下时不适宜做因子分析。

- Bartlett's 球形检验

用于检验相关矩阵是否是单位矩阵，即各变量是否各自独立。如果结论为不拒绝假设，则说明这些变量可能各自独立提供一些信息，之间恐怕没有什么关系。

12.1.5 常用概念和分析步骤

主成分分析和因子分析中常用的主要概念如下。

- 因子负荷

即因子分析式中各因子的系数值，用于反映因子和各个变量间的密切程度。当各公因子间不相关时，因子负荷值就等于因子与变量的相关系数。它的绝对值越大，说明该因子对当前变量的影响程度越大。

- 公因子方差比（Communalities）

是指提取公因子后，各变量中信息分别被提取出的比例，或者说原变量的方差中由公因子决定的比例。公因子方差比在 0~1 之间，取值越大，说明该变量能被因子说明的程度越高，如果各因子间完全独立，则公因子方差比和因子负荷实际上是等价的。

- 特征值（Eigenvalue）

特征值可以被看成是主成分分析影响力度的指标，代表引入该因子或主成分后可以解释平均多少原始变量的信息。

主成分分析和因子分析的标准步骤如下：

01 根据具体问题，判断是否需要进行因子分析，并采用 KMO 检验及 Bartlett's 球形检验来判断数据是否符合分析要求；

02 进行分析，按一定标准确定提取的因子数目；

03 如果进行的是主成分分析，则将主成分存在的新变量用于继续分析，步骤到此结束；

04 如果进行的是因子分析，则考察因子的可解释性，并在必要时进行因子旋转，以寻求最佳解释方式；

05 如有必要，可计算出因子得分等中间指标供进一步分析使用。

12.1.6 主成分和公因子数量的确定

主成分分析希望用尽可能少的主成分包含原来尽可能多的信息，那么如何确定需要保

留的主成分数量？可以遵循以下几点原则。

- 主成分的累积贡献率

一般来说，提取主成分的累积贡献率达到 80%~85%以上就比较满意了，可以由此决定需要提取多少个主成分。

- 特征值

特征值在某种程度上可以被看成是表示主成分影响力度大小的指标，如果特征值小于 1，说明该主成分的解释力度还不如直接引入一个原变量的平均解释力度大。因此一般可以用特征值大于 1 作为纳入标准。

- 综合判断

大量的实际情况表明，如果根据累积贡献率来确定主成分数往往较多，而用特征值来确定又往往较低，很多时候应当将两者结合起来，以综合确定合适的数量。

12.1.7　主成分和因子分析的联系与区别

- 两者都是在多个原始变量中通过它们之间的内部相关性来获得新的变量（主成分变量或公因子变量），达到既减少分析指标个数，又能概括原始指标主要信息的目的。但它们各有其特点：主成分分析是将 m 个原始变量提取 k（$k \leqslant m$）个互不相关的主成分；因子分析是提取 k（$k \leqslant m$）个支配原始变量的公因子和 1 个特殊因子，各公因子之间可以相关或互不相关。
- 提取公因子的方法主要有主成分法和公因子法，若采用主成分法，则主成分分析和因子分析基本等价，该法主要从解释变量的总方差角度，尽量使变量的方差被主成分解释，即主成分法倾向得到更大的共性方差；而公因子法主要是从解释变量的相关性角度，尽量使变量的相关程度能被公因子解释，当因子分析的目的重在确定结构时则会用到该法。
- 因子分析提取的公因子比主成分分析提取的主成分更具有可解释性。主成分分析不考虑观察变量的度量误差，直接用观察变量的某种线性组合来表示一个综合变量；而因子分析的潜在变量则校正了观察变量的度量误差，且它还可进行因子旋转，使潜在因子的实际意义更明确，分析结论更真实。
- 两者分析的实质及重点不同。主成分的数学模型为 $Z = BX$，即主成分 Z 为原始变量 X 的线性组合；因子分析的数学模型为 $X = BF + \varepsilon$，即原始变量 X 为公因子 F 与特殊因子 ε 的线性组合。因此，主成分分析主要是综合原始变量的信息，而因子分析中是解释原始变量之间的关系。主成分分析实质上是线性变换，无假设检验，而因子分析是统计模型，某些因子模型（如 ML 估计）是可以得到假设检验的。
- 两者的 SPSS 操作都是通过“ Analyze→Data Reduction→Factor ”过程实现，但主成分分析主要用“ Descriptives、Extraction、Scores ”对话框，而因子分析除了使用这些对话框外，还可使用“ Rotation ”对话框进行因子旋转。

12.2 因子分析

因子分析是多元统计分析的一个重要分支。其主要目的是运用对诸多变量的相关性研究，即可以用假设的少数几个变量来表示原来变量的主要信息，以便浓缩数据（Data Reduction）。

12.2.1 因子分析的数学模型

因子分析过程是将多个变量表示为较少的因子，下面我们利用数学模型表示因子分析的主要过程。

设有 n 个原始变量，表示 $x_1, x_2, ..., x_n$，根据因子分析的要求，假设这些变量已经标准化（均值为 0，标准差 1），假设 n 个变量可以由 k 的因子 $f_1, f_2, ..., f_k$ 表示为线性组合，即：

$$\begin{cases} x_1 = a_{11} f_1 + a_{12} f_2 + ... + a_{1k} f_k + \varepsilon_1 \\ x_2 = a_{21} f_1 + a_{22} f_2 + ... + a_{2k} f_k + \varepsilon_2 \\ ... \\ x_n = a_{n1} f_1 + a_{n2} f_2 + ... + a_{nk} f_k + \varepsilon_n \end{cases} \tag{12.1}$$

上式为因子分析的数学模型，如果利用矩阵形式则表示为 $X = AF + \varepsilon$。其中 X 为可观测的 n 维变量向量，它的每一个分量表示一个指标或变量；F 称为因子向量，每一个分量表示一个因子，由于它们出现在每个原始变量的线性表达式中，所以又称为公共因子；矩阵 A 为因子负荷矩阵，其元素 a_{ij} 称为因子负荷；ε 称为特殊因子，表示原始变量中不能由因子解释的部分，均值为 0。

因子分析的基本思想是通过对变量的相关系数矩阵的内部结构进行分析的，从中找出少数几个能够控制原始变量的因子 $f_1, f_2, ..., f_k$，选取公共因子的原则是尽可能包含更多的原始变量信息，建立因子分析模型，利用公共因子 $f_1, f_2, ..., f_k$ 再现原始变量之间的相关系数，达到简化变量、降低变量维数和对原始变量再解释及命名的目的。

12.2.2 因子分析的基本方法

在理论分析和具体 SPSS 操作方面，因子分析过程需经过如下几个重要步骤。

01 因子提取。

通过分析原始变量之间的相互关系，从中提取出数量较少的因子。提取方法是利用样本数据得到因子负荷矩阵。求解因子负荷矩阵的方法有很多，其中最常用的是主成分分析法、主轴因子法等，利用因子负荷矩阵求解变量相关矩阵的特征值，根据特征值的大小确定因子数量。

02 因子旋转。

因子分析的一个重要目的是对原始数据进行综合评价。利用因子提取方法得到的结果虽然保证了因子之间的正交性，也就是因子之间不相关，但因子对变量的解释能力较弱，

不容易解释和命名。这时，可以通过对因子模型的旋转变换，使公共因子的负荷系数更接近 1 或更接近 0，通过这种方法得到的公共因子对变量的命名和解释将变得更加容易。

进行正交变换可以保证变换后各因子仍正交，但如果经过正交变换后对公共因子仍然不容易解释，也可以对因子进行斜交旋转变换，可能得到比较容易解释的因子。

03 计算因子得分。

利用因子表示原始变量，需要知道因子和原始变量之间的线性关系。为此，需要计算因子得分，作为进一步分析奠定基础。计算因子得分的方法主要有回归法、巴特利特法和 Anderson-Rubin 法。关于这些方法的定义，可以参考操作过程中对相关选项的解释。

12.2.3　因子分析的实例

下面通过实例来向读者介绍用 SPSS 进行因子分析的具体操作方法。

【示例 12.1】用探索性因子分析方法探讨综合评价体系

为评价医院的医疗工作质量，某研究者收集了近三年的门诊人次、出院人数、病床利用率等 9 个变量，具体数据如表 12.1 所示。试用因子分析方法探讨其综合评价体系（数据文件为 12.1.sav）。

表 12.1　某医院近 3 年医疗工作质量变量数据

年月	门诊人次（万）	出院人数	病床利用率（%）	病床周转次数	平均住院天数	治愈好转率（%）	病死率（%）	诊断符合率（%）	抢救成功率（%）
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
01-01-06	4.34	389	99.06	1.23	25.46	93.15	3.56	97.51	61.66
01-02-06	3.45	271	88.28	0.85	23.55	94.31	2.44	97.94	73.33
01-03-06	4.38	385	103.97	1.21	26.54	92.53	4.02	98.48	76.79
01-04-06	4.18	377	99.48	1.19	26.89	93.86	2.92	99.41	63.16
01-05-06	4.32	378	102.01	1.19	27.63	93.18	1.99	99.71	80
01-06-06	4.13	349	97.55	1.1	27.34	90.63	4.38	99.03	63.16
01-07-06	4.57	361	91.66	1.14	24.89	90.6	2.73	99.69	73.53
01-08-06	4.31	209	62.18	0.52	31.74	91.67	3.65	99.48	61.11
01-09-06	4.06	425	83.27	0.93	26.56	93.81	3.09	99.48	70.73
01-10-06	4.43	458	92.39	0.95	24.26	91.12	4.21	99.76	79.07
01-11-06	4.13	496	95.43	1.03	28.75	93.43	3.5	99.1	80.49
01-12-06	4.1	514	92.99	1.07	26.31	93.24	4.22	100	78.95
02-01-06	4.11	490	80.9	0.97	26.9	93.68	4.97	99.77	80.53
02-02-06	3.53	344	79.66	0.68	31.87	94.77	3.59	100	81.97
02-03-06	4.16	508	90.98	1.01	29.43	95.75	2.77	98.72	62.86
02-04-06	4.17	545	92.98	1.08	26.92	94.89	3.14	99.41	82.35
02-05-06	4.16	507	95.1	1.01	25.82	94.41	2.8	99.35	60.61
02-06-06	4.86	540	93.17	1.07	27.59	93.47	2.77	99.8	70.21
02-07-06	5.06	552	84.38	1.1	27.56	95.15	3.1	98.63	69.23
02-08-06	4.03	453	72.69	0.9	26.03	91.94	4.5	99.05	60.42
02-09-06	4.15	529	86.53	1.05	22.4	91.52	3.84	98.58	68.42
02-10-06	3.94	515	91.01	1.02	25.44	94.88	2.56	99.36	73.91

续表

年月 x_0	门诊人次（万） x_1	出院人数 x_2	病床利用率（%） x_3	病床周转次数 x_4	平均住院天数 x_5	治愈好转率（%） x_6	病死率（%） x_7	诊断符合率（%） x_8	抢救成功率（%） x_9
02-11-06	4.12	552	89.14	1.1	25.7	92.65	3.87	95.52	66.67
02-12-06	4.42	597	90.18	1.18	26.94	93.03	3.76	99.28	73.81
03-01-06	3.05	437	78.81	0.87	23.05	94.46	4.03	96.22	87.1
03-02-06	3.94	477	87.34	0.95	26.78	91.78	4.57	94.28	87.34
03-03-06	4.14	638	88.57	1.27	26.53	95.16	1.67	94.5	91.67
03-04-06	3.87	583	89.82	1.16	22.66	93.43	3.55	94.49	89.07
03-05-06	4.08	552	90.19	1.1	22.53	90.36	3.47	97.88	87.14
03-06-06	4.14	551	90.81	1.09	23.06	91.65	2.47	97.72	87.13
03-07-06	4.04	574	81.36	1.14	26.65	93.74	1.61	98.2	93.02
03-08-06	3.93	515	76.87	1.02	23.88	93.82	3.09	95.46	88.37
03-09-06	3.9	555	80.58	1.1	23.08	94.38	2.06	96.82	91.79
03-10-06	3.62	554	87.21	1.1	22.5	92.43	3.22	97.16	87.77
03-11-06	3.75	586	90.31	1.12	23.73	92.47	2.07	97.14	93.89
03-12-06	3.77	627	86.47	1.24	23.22	91.17	3.4	98.98	89.8

01 导入数据。

依次单击菜单“文件→打开→数据文件”命令，或者直接单击 SPSS 窗口 图标，打开如图 12.2 所示的导入数据对话框。

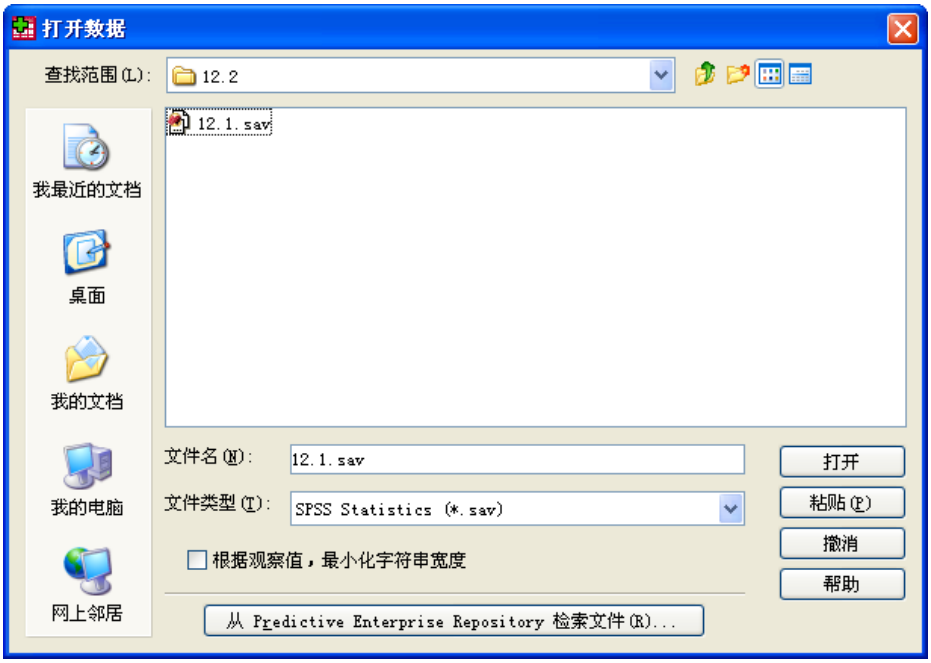


图 12.2　导入数据

02 因子分析。

依次单击菜单“ 分析→降维→因子分析 ”命令，如图 12.3 所示。打开图 12.4 所示的“ 因子分析 ”主对话框。

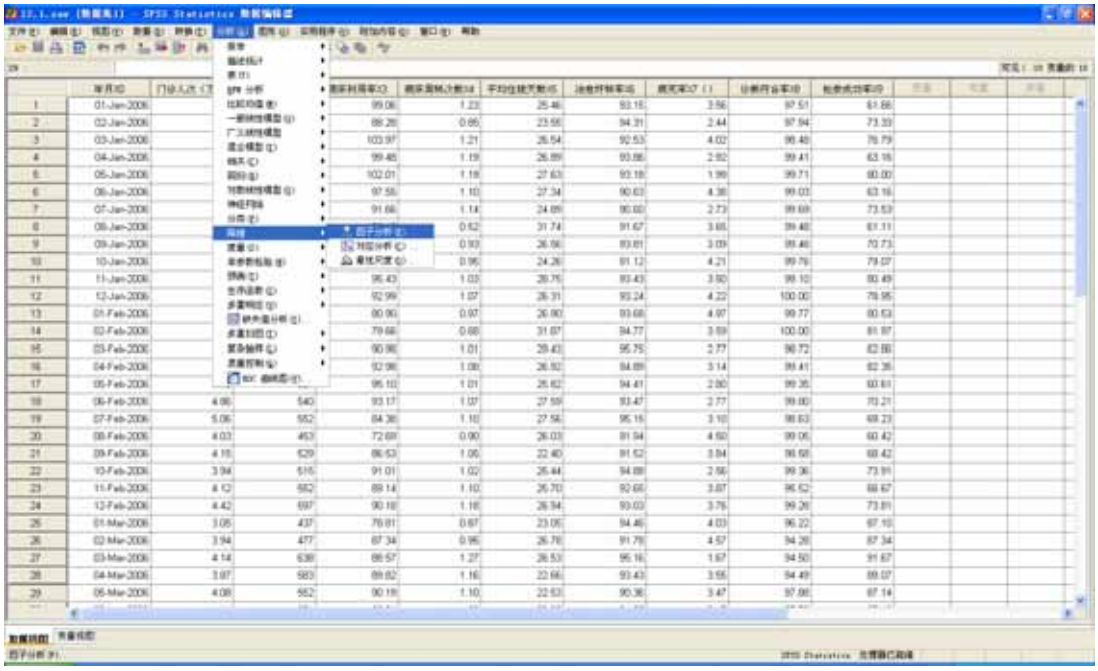


图 12.3　因子分析菜单



(a) 选入变量前　　(b) 选入变量后

图 12.4 “ 因子分析 ”主对话框

在图 12.4a 所示的对话框中选中左边的变量，单击 按钮，可以将其选入到右边的列表框中（如图 12.4 b 所示）。单击“ 描述 ”按钮，弹出“ 因子分析：描述统计 ”对话框，如图 12.5 所示，在“ 统计量 ”选项组中选取“ 原始分析结果 ”。设置完毕后，单击“ 继续 ”按钮，确认操作，返回主对话框；也可以单击“ 取消 ”按钮，删除操作；单击“ 帮助 ”按钮，进行在线帮助操作。

在图 12.4 所示的对话框中单击“ 抽取 ”按钮，得到如图 12.6 所示的“ 因子分析：抽取 ”对话框。选择“ 方法 ”为“ 主成分 ”；在“ 分析 ”选项组选择“ 相关性矩阵 ”；在“ 输出 ”选项组选择“ 未旋转的因子解 ”；在“ 提取 ”选项组将“ 因子的固定数量：” 设置为 4；将“ 最大收敛性迭代次数：” 设置为 25。

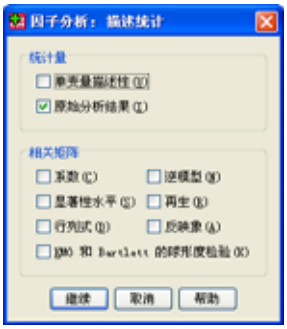


图 12.5 “ 因子分析：描述 ”对话框

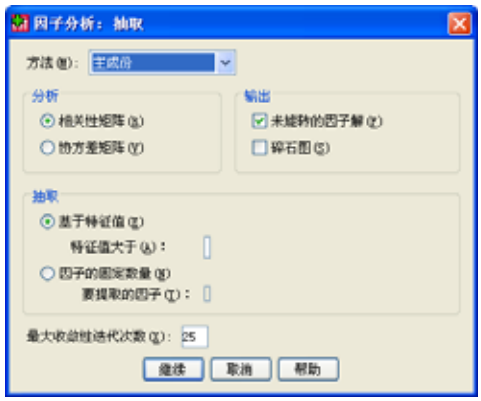


图 12.6 “ 因子分析：抽取 ”对话框

在图 12.4 所示的对话框中单击“ 旋转 ”按钮，得到如图 12.7 所示的“ 因子分析：旋转 ”对话框。在“ 方法 ”选项组选择“ 最大四次方值法 ”；在“ 输出 ”选项组选择“ 旋转解 ”；将“ 最大收敛性迭代次数：” 设置为 25。

在图 12.4 所示对话框中单击“ 得分 ”按钮，得到如图 12.8 所示的“ 因子分析：因子得分 ”对话框。选择“ 保存为新变量 ”和“ 显示因子得分系数矩阵 ”；在“ 方法 ”选项组选择“ 回归 ”。

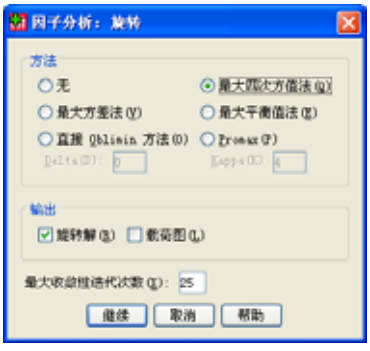


图 12.7 “ 因子分析：旋转 ”对话框

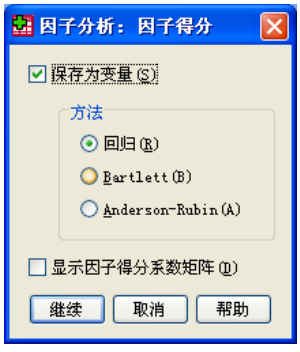


图 12.8 “ 因子分析：因子得分 ”对话框

在“ 因子分析 ”主对话框（如图 12.4 所示）中，单击“ 确定 ”按钮，执行操作。

03 结果解释。

- 主成分信息（如表 12.2 所示）

表 12.2　主成分信息

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.822	31.357	31.357	2.8	31.357	31.357	2.646	29.404	29.404
2	1.992	22.138	53.495	22	22.138	53.495	1.847	20.527	49.931
3	1.448	16.091	69.585	1.9	16.091	69.585	1.471	16.340	66.271
4	.784	8.709	78.294	92	8.709	78.294	1.082	12.023	78.294
5	.668	7.424	85.718	1.4					
6	.537	5.965	91.683	48					
7	.454	5.047	96.730	.78					
8	.175	1.942	98.672	4					
9	.119	1.328	100.000						

Extraction Method：Principal Component Analysis.

表 12.2 中显示前 3 个主成分的特征值大于 1，但它们的累积贡献率仅为 69.585%。即约 70.3%的总方差可以由以下四个潜在因子解释。

- 变量的共同度（如表 12.3 所示）

表 12.3　变量的共同度

Communalities

	Initial	Extraction
门诊人次（万）X1	1.000	.880
出院人数 X2	1.000	.873
病床利用率 X3（%）	1.000	.873
病床周转次数 X4	1.000	.917
平均住院天数 X5	1.000	.767
治愈好转率 X6（%）	1.000	.796
病死率 X7（%）	1.000	.683
诊断符合率 X8（%）	1.000	.573
抢救成功率 X9（%）	1.000	.684

Extraction Method：Principal Component Analysis.

结果显示，每一个变量的共性方差均在 0.5 以上，且大多数接近或超过 0.7，这说明 4 个公因子能够较好地反映客观原变量的大部分信息。

- 旋转前的因子负荷矩阵（如表 12.4 所示）

表 12.4　旋转前的因子负荷矩阵

Component Matrix[a]

	Component			
	1	2	3	4
门诊人次（万）X1	-.260	.769	.009	.469
出院人数 X2	.764	.133	.090	.513
病床利用率 X3（%）	.239	.778	-.085	-.452
病床周转次数 X4	.684	.666	.070	-.024
平均住病天数 X5	-.724	.119	.441	.185
治愈好转率 X6（%）	.039	-.070	.889	-.021
病死率 X7（%）	-.406	-.163	-.663	.230
诊断符合率 X8（%）	-.637	.397	.039	-.090
抢救成功率 X9（%）	.740	-.362	.057	.034

Extraction Method：Principal Component Analysis.

a．4 component extracted.

根据 0.5 原则，因子 1 在多数原始变量上有较大的负荷，因子 2 在 x_1（门诊次数）、x_3（病床利用率）和 x_4（病床周转次数）变量上有较大负荷，因子 3 在 x_6（治愈好转率）、x_7（病死率）变量上有较大负荷，因子 4 在 x_2（出院人数）变量上有较大负荷。因此说明，除可初步认定因子 1 反映综合情况、因子 3 反映医疗水平情况外，其他两个因子的意义不明显。

- 正交旋转矩阵（如表 12.5 所示）

表 12.5　正交旋转矩阵

Component Transformation Matrix

Component	1	2	3	4
1	-.899	.387	.153	.138
2	.413	.786	.027	.460
3	.118	-.140	.980	.076
4	-.086	-.462	-.124	.874

Extraction Method：Principal Component Analysis.

Rotation Method：Quartimax with Kaiser Normalization.

a．Rotation converged in 5 iterations.

这是通过四次方最大旋转得到的正交变换矩阵。

- 旋转后的因子负荷矩阵（如表 12.6 所示）

表 12.6　旋转后的因子负荷矩阵

Rotated Component Matrix[a]

	Component			
	1	2	3	4
门诊人次（万）X1	.512	.285	-.068	.729

续表

	Component			
	1	2	3	4
出院人数 X2	-.666	.151	.146	.621
病床利用率 X3（%）	.135	.924	.030	-.010
病床周转次数 X4	-.346	.809	.057	.374
平均住院天数 X5	.736	-.334	.302	.150
治愈好转率 X6（%）	.044	-.155	.877	.023
病死率 X7（%）	.199	-.298	-.744	.019
诊断符合率 X8（%）	.749	.102	-.038	.020
抢救成功率 X9（%）	-.811	-.022	.156	-.030

Extraction Method：Principal Component Analysis.

Rotation Method：Quartimax with Kaiser Normalization.

a．Rotation converged in 5 iterations

通过四次方最大旋转后，得到了 9 个变量在 4 个因子上的新的因子负荷。结果显示，因子 1 支配的变量有 x_1（门诊次数）、x_2（出院人数）、x_5（平均住院天数）、x_8（诊断符合率）和 x_9（抢救成功率），因子 2 支配的变量有 x_3（病床利用率）和 x_4（病床周转次数），因子 3 支配的变量有 x_6（治愈好转率）和 x_7（病死率），因子 4 支配的变量有 x_2（出院人数）。故可以认为，因子 1 反映医院医疗工作质量各方面的情况，称为综合因子，因子 2 反映病床利用情况，称为病床利用因子，因子 3 反映医疗水平，称为水平因子，因子 4 反映就诊病人数量，称为数量因子。与旋转前的因子负荷矩阵相比较，说明该旋转对因子负荷起到了明显的分离作用，使各因子具有较明显的专业意义。

通过探索性因子分析，从这 9 个医院医疗工作质量变量中找出了 4 个潜在因子，它们为：综合因子、病床利用因子、水平因子和数量因子。它们之间没有交叉支配，即每个变量只受一个潜在因子影响，且没有单变量潜在因子出现，即一个潜在因子至少支配两个变量。

图 12.9 所示为将通过旋转后计算出的每条记录的 4 个因子得分作为新变量自动存储到原始数据文件中。FAC_1 为第 1 因子得分，FAC_2 为第 2 因子得分，FAC_3 为第 3 因子得分，FAC_4 为第 4 因子得分，根据以上分析，可了解各个观察对象的潜在本质。

12.3　主成分分析

主成分分析的主要目的是希望用较少的变量去解释原来资料中的大部分变异，将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量。通常是选出比原始变量个数少、能解释大部分资料中的变异的几个新变量，即所谓主成分，并用以解释资料的综合性指标。由此可见，主成分分析实际上是一种降维方法。

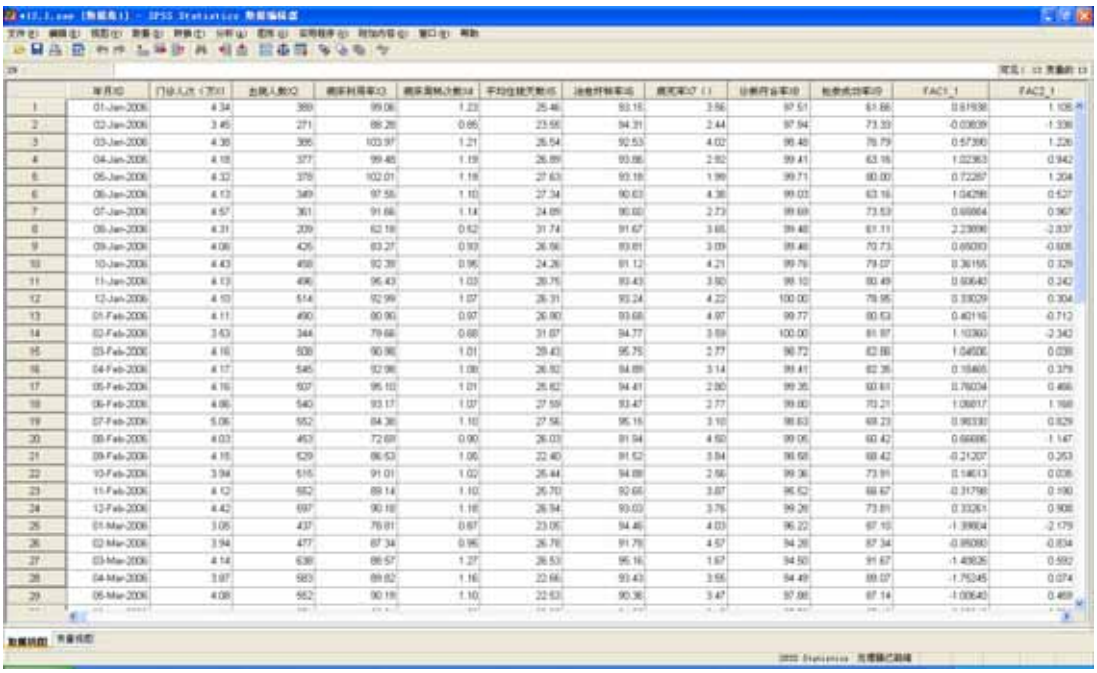


图 12.9　存储数据文件

12.3.1　主成分分析的数学模型

主成分分析旨在利用降维的思想，把多指标转化为少数几个综合指标。

设有 n 个原始变量，表示 $x_1, x_2, ..., x_n$ ， $Y_1, Y_2, ..., Y_n$ 表示主成分后的变量， R 为变量之间的相关系数矩阵，根据主成分分析的要求有：

$$\begin{cases} Y_1 = R_{11}x_1 + R_{12}x_2 + ... + R_{1k}x_k \\ Y_2 = R_{21}x_1 + R_{22}x_2 + ... + R_{2k}x_k \\ ... \\ Y_n = R_{n1}x_1 + R_{n2}x_2 + ... + R_{nk}x_k \end{cases} \tag{12.2}$$

假设头两个成分特征值对应的方差累积占了总方差的 81.142%，累计方差贡献率为 81.142%。后面的特征值的贡献越来越少。一般我们取累计方差贡献率达到 85%左右的前 k 个主成分就可以了，因为它们已经代表了绝大部分的信息。

12.3.2　主成分分析的基本方法

在理论分析和具体 SPSS 操作方面，主成分分析过程需经过如下几个重要步骤：

01 将原始数据标准化，以消除量纲的影响；

02 建立变量之间的相关系数矩阵 R；

03 求 R 的特征值和特征向量；

04 写出主成分并进行分析。

SPSS 中选取主成分的方法有两种：一种是根据特征根大于等于 1 来选取；另一种是根据用户直接规定主成分的个数来选取。

特征值的贡献还可以从 SPSS 的所谓碎石图中看出。

可以把第一和第二主成分的点画出一个二维图，以直观地显示它们是如何解释原来的变量的。

12.3.3　主成分分析的实例

【示例 12.2】用主成分分析法解决自变量的多重共线性问题。

进行多重线性回归分析时，经常碰到自变量之间强相关的问题，即多重共线问题。主成分分析法则是解决这类问题的好办法，可通过主成分回归来求回归系数。主成分回归是将原自变量的主成分代替原自变量进行回归分析，主成分分析保留了原变量的绝大部分信息，又有主成分之间互不相关的特点。

主成分回归的具体步骤如下：

01 采用多重回归分析，进行共线性诊断；

02 进行主成分分析确定所需主成分数；

03 进行主成分回归分析。

按确定的主成分数量，将排在前面的主成分代替原自变量进行多重回归分析，得到标准化自变量与因变量之间的回归模型；然后将标准化自变量还原为原自变量，得到原自变量与因变量的回归模型。下面通过实例说明。

某研究者收集了 13 名儿童的心象面积数据，如表 12.7 所示。其中包括：性别（x_1：男=1，女=2）、年龄（x_2：月）、身高（x_3：cm）、体重（x_4：kg）、胸围（x_5：cm）和心象面积（y：cm^2），试分析心象面积与性别、年龄、身高、体重和胸围之间的关系。

表 12.7　13 名儿童心象面积研究数据

ID x_0	性别 x_1	年龄（月） x_2	身高（cm） x_3	体重（kg） x_4	胸围（cm） x_5	心象面积（cm^2） y
1	1	32	95.5	14.0	53.5	49.64
2	1	35	92.0	13.0	52.0	41.46
3	1	33	89.0	12.5	53.5	35.81
4	1	176	168.0	53.5	82.0	100.14
5	1	96	117.0	19.7	56.0	67.20
6	1	96	113.0	18.1	55.0	60.00
7	1	96	122.0	21.6	57.3	58.00
8	2	30	91.0	11.0	48.0	35.39
9	2	33	91.0	11.5	47.0	44.98
10	2	33	91.0	12.5	50.0	29.51
11	2	176	156.0	55.0	83.0	94.66
12	2	178	163.0	54.0	79.0	87.42
13	2	84	130.0	25.0	58.0	62.00

1．采用多重回归分析，进行共线性诊断。

依次单击菜单“ 分析→回归分析→线性 ”命令，打开线性回归分析主界面，如图 12.10 所示。

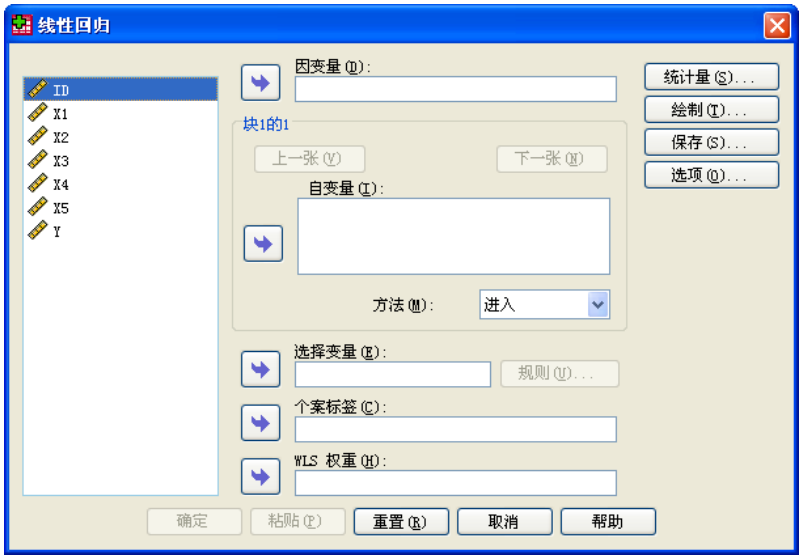


图 12.10　线性回归分析

将变量 Y 选入“ 因变量 ”列表框，将变量 x_1 至 x_5 选入“ 自变量 ”，如图 12.11 所示。

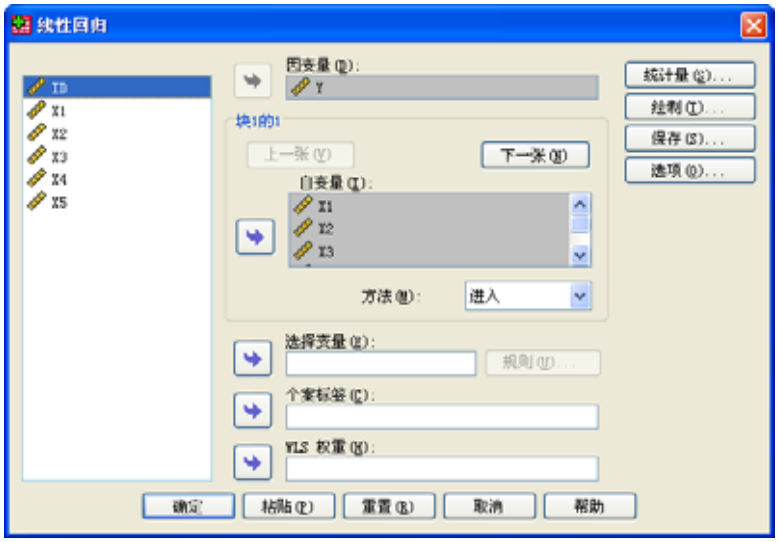


图 12.11　选入变量

单击“ 统计量 ”按钮，并在弹出的“ 线性回归：统计量 ”对话框中选取“ 估计、模型拟合度、描述性、共线性诊断和误差条图的表征 ”，以便获取共线性诊断的有关结果，如

图 12.12 所示。

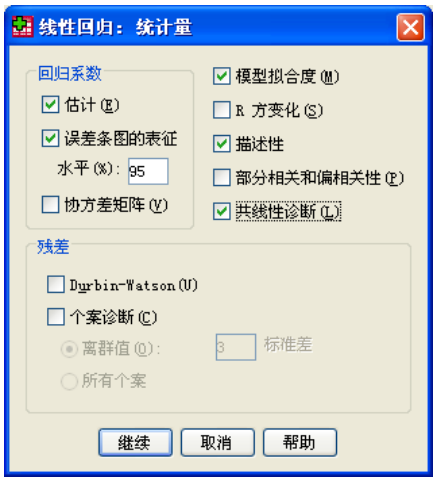


图 12.12 “ 线性回归：统计量 ”对话框

多重回归分析及共线性诊断的有关结果如下。

- 模型总体的假设检验结果

表 12.8　模型总体的假设检验结果

Model Summary

Model	R	R Square	Adjusted R Square	Std Error of the Estimate
1	.976[a]	.953	.920	6.53044

a．Predictors：(Constant)，X5，X1，X2，X3，X4

a．假设检验结果

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6108.187	5	1221.637	28.646	.000[a]
	Residual	298.526	7	42.647		
	Total	6406.713	12			

a．Predictors：(constant)，X5，X1，X2，X3，X4

b．Dependent Variable：Y

b．方差分析表

注释：表 12.8 为模型总体的假设检验结果和方差分析表。在（a）假设检验结果中，R 为相关系数，R Sequare 为判定系数，Aajusted R Square 为调整判定系数，Std.Error of the Estimate 为估计值的标准误差。在（b）方差分析表中，Sum of Squares 为偏差平方和，df 为样本的自由度，Mean Square 为偏差均方值，

F 为 *F* 检验统计量的观察值，Sig.为自由度的熵。Regression 表示回归统计，Residual 为剩余统计，Total 为两者之和。

从表中可以看出，模型总体拟合很好（R^2=0.953），方差分析表显示结果有统计学意义（*P*=0.000）。

- 参数估计及其假设检验结果

表 12.9　参数估计及其假设检验结果

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
		B	Std.Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	（Constant）	54.582	124.273		.439	.674	-239.276	348.440		
	X1	-7.763	8.066	-174	-.962	.368	-26.836	11.311	.203	4.929
	X2	.121	.180	.309	.672	.523	-.304	.546	.031	31.803
	X3	.290	.418	.368	.693	.510	-.698	1.278	.024	42.363
	X4	1.121	2.256	.840	.497	.634	-4.212	6.455	.002	428.935
	X5	-.941	2.332	-.524	-404	.699	-6.456	4.574	.004	253.206

a．Dependent Variable：Y

注释：表 12.9 中 Unstandardized Coefficients 和 Standardized Coefficients 分别表示非标准线性回归系数和标准线性回归系数，B 为回归系数，Std.Error 为标准误差，Beta 为标准化回归系数，t 为回归系数 T 检验的 *t* 统计量观察值，Sig.为自由度的熵，Confidence Interval for Mean 为置信区间（Lower Bound 表示置信区间的下限，Upper Bound 表示置信区间的上限），Collinearity Statistics 为共线性统计（Tolerance 为公差，VIF 为方差扩大化因子）。Constant 表示常数项系数。

从表中可以看出，尽管模型总体拟合较好，有统计学意义，但参数估计结果显示各偏回归系数均无统计学意义，说明自变量存在共线性。

- 共线性诊断

表 12.10　共线性诊断结果

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	X1	X2	X3	X4	X5
1	1	5.581	1.000	.00	.00	.00	.00	.00	.00
	2	.335	4.084	.00	.01	.01	.00	.00	.00
	3	.071	8.870	.00	.17	.00	.00	.00	.00
	4	.012	22.024	.00	.04	.39	.00	.02	.00
	5	.001	62.706	.01	.02	.59	.65	.00	.01
	6	8.11E-005	262.325	.99	.76	.02	.34	.97	.98

a．Dependent Variable：Y

表 12.10 显示自变量存在严重的共线性（条件指数 Φ =262.325），常数项（Constant）、x_4 、 x_5 的 VP（Variance Proportions）值均很大，分别是 0.99、0.97 和 0.98，因此，自变量 x_4 和 x_5 与常数项极度相关。因此我们需要采用主成分回归分析。

- 描述性统计量

表 12.11　描述性统计量

Descriptive Statistics

	Mean	Std.Deviation	N
Y	58.9392	23.10612	13
X1	1.46	.519	13
X2	84.46	59.173	13
X3	116.808	29.3638	13
X4	24.723	17.3091	13
X5	59.562	12.8624	13

表 12.11 显示了原自变量的均值、标准差和例数信息。

- 相关系数矩阵

表 12.12　相关系数矩阵

Correlations

		Y	X1	X2	X3	X4	X5
Pearson Correlation	Y	1.000	.002	.967	.965	.940	.935
	X1	.002	1.000	.074	.116	.192	.095
	X2	.967	.074	.981	.981	.960	.949
	X3	.965	.116	1.000	1.000	.966	.950
	X4	.940	.192	.966	.966	1.000	.991
	X5	.935	.195	.950	.950	.991	1.000s
Sig.(1-tailed)	Y	.	.497	.000	.000	.000	.000
	X1	.497	.	.405	.353	.265	.378
	X2	.000	.405	.	.000	.000	.000
	X3	.000	.353	.000	.	.000	.000
	X4	.000	.265	.000	.000	.	.000
	X5	.000	.378	.000	.000	.000	.
N	Y	13	13	13	13	13	13
	X1	13	13	13	13	13	13
	X2	13	13	13	13	13	13
	X3	13	13	13	13	13	13
	X4	13	13	13	13	13	13
	X5	13	13	13	13	13	13

注释：表 12.12 中，Y 为因变量，X1、X2、X3、X4 和 X5 为自变量。Pearson Correlation 为泊松相关系数矩阵，Sig.（1-tailed）为自由度的熵，N 为样本统计量的个案数。

表 12.12 显示了原自变量的相关系数矩阵及相应 *P* 值。

2．进行主成分分析确定所需主成分数。

依次单击菜单“ 分析→降维→因子分析 ”命令，打开“ 主成分分析 ”对话框。将变量选入右边的列表框，如图 12.13 所示。

单击“ 描述 ”按钮，打开图 12.14 所示的“ 因子分析：描述统计 ”对话框。在“ 统计量 ”选项组中选择“ 单变量描述性 ”和“ 原始分析结果 ”选项；在“ 相关矩阵 ”选项组中选择“ 系数 ”和“ 显著性水平 ”选项。



图 12.13 “ 主成分分析 ”窗口

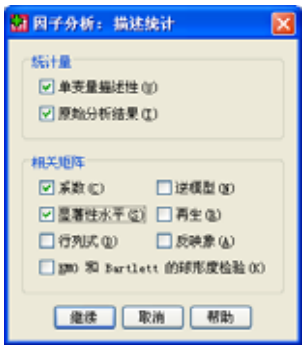


图 12.14 “ 因子分析：描述统计 ”对话框

单击“ 抽取 ”按钮，打开如图 12.15 所示的“ 因子分析：抽取 ”对话框。将“ 方法 ”设置为“ 主成分分析法 ”；在“ 分析 ”选项组中选择“ 相关性矩阵 ”；在“ 输出 ”选项组中选择“ 未旋转的因子解 ”；将“ 因子的固定数量：”设置为 5；将“ 最大收敛性迭代次数：”设置为 25。

单击“ 得分 ”按钮得到如图 12.16 所示的“ 因子分析：因子得分 ”对话框。选中“ 保存为变量 ”和“ 显示因子得分系数矩阵 ”；在“ 方法 ”选项组中选择“ 回归 ”。

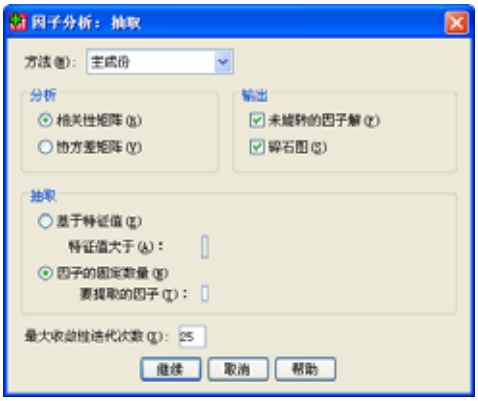


图 12.15 “ 因子分析：抽取 ”对话框



图 12.16 “ 因子分析：得分 ”对话框

主成分分析的有关结果如下。

- 主成分的统计信息。

表 12.13　主成分的统计信息

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.918	78.366	78.366	3.918	78.366	78.366
2	.989	19.770	98.136	.989	19.770	98.136
3	.073	1.462	99.598	.073	1.462	99.598
4	.019	.373	99.971	.019	.373	99.971
5	.001	.029	100.000	.001	.029	100.000

Extraction Method：Principal Component Analysis.

表 12.13 显示了主成分的统计信息。包括特征值由大到小的次序排列，各主成分的贡献率及累积贡献率。第一主成分的特征值为 3.918，它解释了 5 个原变量的总方差的 78.366% ;第二主成分的特征值为 0.989，接近 1，它解释了 5 个原变量的总方差的 19.770%。前两个特征值的累积贡献率为 98.136%，即前两个主成分包含了原有 5 个变量的 98.136% 的信息，所以本例可以取前两个主成分来代替原有的 5 个指标变量。

- 碎石图。

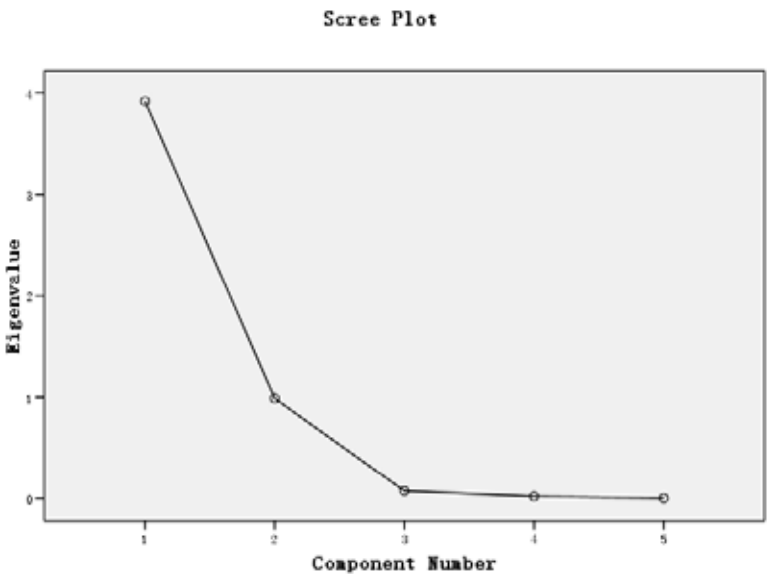


图 12.17　碎石图

图 12.17 显示前两个主成分的特征值接近 1 以上，进一步说明取前两个主成分。

- 因子得分系数矩阵

表 12.14　因子得分系数矩阵

Component Score Coefficient Matrix

	Component				
	1	2	3	4	5
X1	.041	.998	.142	.252	1.961
X2	.251	-.089	1.839	5.160	1.315
X3	.252	-.045	1.785	-5.176	3.511
X4	.254	.033	-1.325	-.301	-20.665
X5	.251	-.063	-2.309	.297	15.739

Extraction Method：Principal Component Analysis.

Component Scores.

如表 12.28 所示为因子得分系数矩阵，通过该矩阵可以将所有主成分表示为各个变量的线性组合。本例可以写出两个主成分的表达式如下：

$$\begin{cases} z_1 = 0.041\text{std}x_1 + 0.251\text{std}x_2 + 0.252\text{std}x_3 + 0.254\text{std}x_4 + 0.251\text{std}x_5 \\ z_2 = 0.998\text{std}x_1 - 0.089\text{std}x_2 - 0.045\text{std}x_3 + 0.033\text{std}x_4 - 0.063\text{std}x_5 \end{cases} \tag{12.3}$$

这里 $\text{std}x_i\,(i=1,2,3,4,5)$ 表示标准指标变量。

$$\begin{cases} \text{std}x_1 = (x_1 - 1.46)/0.519 \\ \text{std}x_2 = (x_2 - 84.46)/59.173 \\ \text{std}x_3 = (x_3 - 116.81)/29.364 \\ \text{std}x_4 = (x_4 - 24.72)/17.309 \\ \text{std}x_5 = (x_5 - 59.56)/12.862 \end{cases} \tag{12.4}$$

根据上面公式可以计算出每条记录的第一和第二主成分标准化值，它们与系统自动存储为新变量的主成分结果是一致的。

3．主成分回归分析。

依次单击菜单“分析→回归分析→线性”命令，打开如图 12.10 所示的对话框，调用多重线性回归分析。将 Y 选入“Dependent”；将“FAC1_1，FAC2_1”选入“Independent(s)”。

主成分回归分析的有关结果如下。

- 主成分回归分析模型

表 12.15　主成分回归分析模型结果

Model Summary

Modle	R	R Square	Adjusted R Square	Std.Error of the Estimate	Change Statistice				
					R Square Change	F Change	df1	df2	Sig.F Change
1	.971[a]	.943	.932	6.04470	.943	82.671	2	10	.000

a．Predictors：(Constant),REGR factor score 2 for analysis 1,REGR factor score 1 for analysis 1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6041.329	2	3020.665	82.671	.000[a]
	Residual	365.384	10	36.538		
	Total	6406.713	12			

a．Predictors：(Constant),REGR factor score 2 for analysis 1,REGR

b．Dependent Variable：Y

表 12.15 所示为主成分回归分析的模型拟合情况，结果显示模型较好拟合（R^2=0.943，方差分析 P=0.000）。

- 主成分回归分析的参数估计及其假设检验

表 12.16　主成分回归分析的参数估计及其假设检验结果

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std.Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	58.939	1.676		35.156	.000	55.204	62.675					
	REGR factor score 1 for analysis 1	22.149	1.745	.959	12.693	.000	18.261	26.037	.959	.970	.959	1.000	1.000
	REGR factor score 2 for analysis 1	-3.589	1.745	-.155	-2.057	.000	-7.477	.299	-.155	-.545	-.155	1.000	1.000

a．Dependent Variable：Y

表 12.16 为主成分回归分析的参数估计及其假设检验。结果显示 β_0、β_1 和 β_2 均有统计学意义（P=0.000，0.000，0.067），即 z_1（FAC1_1）和 z_2（FAC2_1）对因变量 Y 有作用，其线性回归方程为：

$$\hat{Y} = 58.939 + 22.149z_1 - 3.589z_2 \tag{12.5}$$

- 将主成分分析结果引入（如表 12.16 所示）

将公式 (12.3) 的 z_1、z_2 代入回归方程 (12.5)，得到因变量 Y 与标准自变量 $\mathrm{std}x_1 \sim \mathrm{std}x_5$ 的线性回归方程：

$$\hat{Y} = 58.951 - 2.683\mathrm{std}x_1 + 5.877\mathrm{std}x_2 + 5.741\mathrm{std}x_3 + 5.505\mathrm{std}x_4 + 5.784\mathrm{std}x_5 \tag{12.6}$$

- 将标准自变量还原为原自变量

将主成分分析中提到的公式（12.4）代入公式（12.5）的回归方程中，得到的因变量 Y 与原自变量 $x_1 \sim x_5$ 的线性回归方程为：

$$\breve{Y} = 0.626 - 5.169x_1 + 0.099x_2 + 0.196x_3 + 0.318x_4 + 0.450x_5 \tag{12.7}$$

在上述分析步骤中，最后两步需要人工计算，其余过程均通过 SPSS 实现。

12.4 本章小结

本章对因子分析和主成分分析的基本概念、基本原理和分析步骤进行了简单的概述，重点讲述了 SPSS 因子分析和主成分分析的基本过程和操作步骤。SPSS 因子分析和主成分分析的操作设置对话框较多，输出结果较为复杂，重点是理解共同度表、负荷矩阵表、总方差解释表和因子碎石图的基本含义。

第13章　可靠性分析

可靠性分析（Reliability Analysis）又称为信度分析，是一种度量综合评价体系是否具有一定的稳定性和可靠性的有效分析方法。例如，在教育学上可衡量教学评价过程受干扰因素所造成的随机误差的大小，可靠性和效度在教育学方面是衡量考试质量的两个重要指标。类似问题在社会生活或经济管理活动中非常普遍，再比如，市民对政府公务员工作满意度的评价，医生对病人的身体状况进行综合评分，挑选出国留学人员要对被选人员的整体素质进行评价等。

综合评价问题在统计学方面可利用可靠性分析方法进行分析。基本方式是做出被评估对象的总体目标，然后将其分解为若干个子目标，它们是总体目标不同方面的体现，是总体特征的部分反映。进一步，再将每个子目标进行量化处理。评估者通过计算被评估对象的总体得分得到最终的评估结果。

上述过程实际上是编制量表的过程。编制量表是否合理决定评价结果的可靠性和效度。SPSS 的可靠性分析的主要作用在于对量表内在可靠性的分析。

在 SPSS 中可以利用“ 分析→度量→可靠性分析 ”命令进行可靠性分析，菜单内容如图 13.1 所示。

本章学习要点：

- 可靠性分析概述；
- SPSS 可靠性分析；
- 可靠性分析的其他问题。

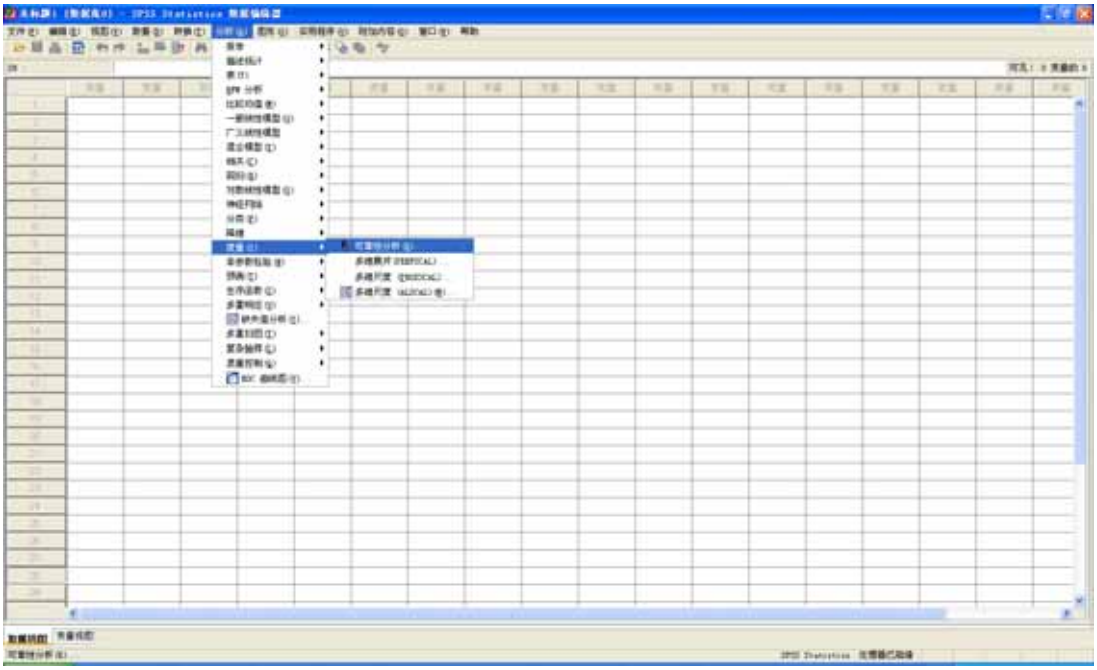


图 13.1　SPSS 可靠性分析模块

13.1　可靠性分析概述

可靠性是指测量的一致性，反映了测验工具所得到的一致性或稳定性，是被测特征真实程度的指标。

13.1.1　可靠性分析的基本原理

现举例说明可靠性的含义。假如要调查某人的文化水平，将文化水平简单地定义为接受正规学校教育的年数。问题是：“你在学校里读了几年书？”接下来记录你的答案。假如能够消除这个人对问题和答案的记忆，提问者会重复问同样的问题并记录下答案。通过考察这个人对同一问题的多次回答，可以判断答案的一致性如何。答案的波动越大，可靠性越低；回答的一致性越好，可靠性越高。

心理测试的理论源自心理学。关于可靠性理论的基本公式是：

$$X_i = \tau_i + e_i \tag{13.1}$$

其中，X_i 是第 i 次测量的得分，e_i 是误差项，τ_i 是关于 X_i 的真实分数。假定真实分数和误差项之间不相关，误差项的均值等于零，即 $COV(\tau_i, e_i) = 0$，$E(e_i) = 0$。按照经典的测量理论，各个测量的误差项是不相关的，测量得分之间的相关是它们真实得分之间的相关造成的。

可靠性定义为：

$$\text{信度} = \frac{VAR(\tau_i)}{VAR(x_i)} \tag{13.2}$$

可靠性是真实分数的方差和实际测量得分的方差的比，它等于实际得分和真实分数的平方相关系数。

一般而言，两次或两个测验的结果越接近，则误差越小，所得的可靠性越高。可靠性是指测验所得到结果的一致性或稳定性，而非测验或量表本身；可靠性值强调的是某一特定类型下的一致性，可靠性系数会因不同时间、不同受试者或不同评分者而出现不同的结果。在统计学分析方面可靠性检验完全依赖于统计方法。

根据被测试者的测试时间和测试内容，可靠性又可分为以下两类。

第一类内在可靠性

对一组问题是否测量同一个概念，同时组成量表题项的一致性程度如何；常用的检测方法是 Cronbach's α 系数。

第二类外在可靠性

对相同的测试者在不同时间测得的结果是否一致，再测可靠性是外在可靠性最常用的检验法。

可靠性指标是对可靠性的一种定量化的描述方式，可靠性指标的量化值称为可靠性系数。可靠性系数越大，表明测量的可靠性就越大。不同研究者对可靠性系数的界限值有不同的看法，一般来说，0.60~0.65 认为不可信；0.65~0.75 认为是最小可接受值；0.70~0.80 认为相当好；0.80~0.90 就是非常好。

因此，一份可靠性系数好的量表或问卷最好在 0.80 以上，0.70~0.80 之间还算是可以接受的范围；分量表最好在 0.70 以上，0.60~0.70 之间可以接受。若分量表的内部一致性系数在 0.60 以下或者总量表的可靠性系数在 0.80 以下，应考虑重新修订量表或增删题项。

可靠性指标多以相关系数来表示，大致可分为三类：稳定系数（跨时间的一致性）、等值系数（跨形式的一致性）和内在一致性系数（跨项目的一致性）。

13.1.2　可靠性分析的基本方法

可靠性分析主要有重测可靠性法、折半可靠性法和 α 可靠性系数法三种。下面简述各种不同方法的基本含义和计算公式。

1．重测可靠性法。

用同样的问卷对同一被测间隔一定的时间重复测试，计算两次测试结果的相关系数。很显然这是稳定系数，即跨时间的一致性。重测可靠性法适用于事实性的问卷，也可用于不易受环境影响的态度、意见式问卷。由于重测可靠性需要对同一样本试测两次，而被测容易受到各种事件、活动的影响，所以间隔时间需要适当。较常用的间隔为两星期或一个月。

2．折半可靠性法。

折半可靠性是指将测量项目按奇偶项分成两半分别记分，测算出两半分数之间的相关

系数，再据此确定整个测量的可靠性系数 R_{xx}。折半可靠性属于内在一致性系数，测量的是两半项目间的一致性。这种方法不适合测量事实性问卷，常用于态度、意见式问卷的可靠性分析。在问卷调查中，态度测量最常见的形式是 5 级李可特量表。

进行折半可靠性分析时，如果量表中含有反意题项，应先将反意题项的得分做逆向处理，以保证各题项得分方向的一致性，然后将全部题项按奇偶或前后分为尽可能相等的两半，计算两者的相关系数 r_{xx}，即半个变量的可靠性系数，最后用皮尔曼—布朗（Spearman-Brown）公式：

$$R_{xx}=\frac{r_{xx}}{1+r_{xx}} \tag{13.3}$$

3．α 可靠性系数法。

Combach α 可靠性系数是目前最常用的可靠性系数。其公式为：

$$\alpha=\frac{k}{k-1}(1-\frac{\sum_{i=1}^{k}VAR(i)}{VAR}) \tag{13.4}$$

其中，k 为量表中评估项目的总数，$VAR(i)$ 为第 i 个项目得分的表内方差，VAR 为全部项目总得分的方差。从公式（13.4）可以看出，Combach α 可靠性系数是量表中各项目得分间的一致性，属于内在一致性系数。这种方法适用于态度、意见式问卷（量表）的可靠性分析。

13.2 SPSS 可靠性分析

SPSS 可靠性分析过程属于“ 度量 ”分析模块中的一个过程，具体操作步骤如下。

- 选择可靠性分析对话框

依次单击菜单“ 分析→度量→可靠性分析 ”命令，打开“ 可靠性分析 ”对话框，如图 13.2 所示。

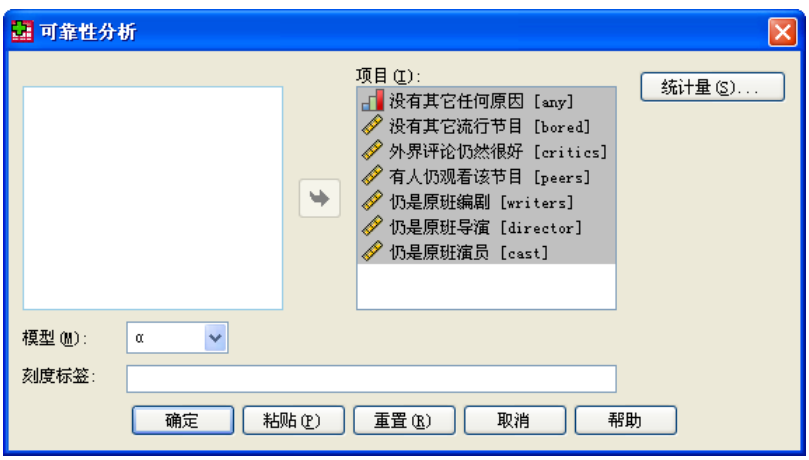


图 13.2 “ 可靠性分析 ”对话框

● 选择评估项目

在左侧变量列表框中选择两个或两个以上变量作为评估项目，单击 按钮，可以将其移动到“项目”列表框中，如图 13.2 所示。

● 选择可靠性分析模型

在图 13.2 所示对话框“模型”下拉列表框中选择可靠性分析模型，SPSS 给出了五种模型：如表 13.1 所示。

表 13.1　可靠性分析模型选项及其含义

选　项	含　义
α	Cornbach α 模型。默认选项。计算用于度量表内部一致性的 Cornbach α 模型
半分	拆半可靠性系数模型。选择该项，计算度量外在可靠性的表间相关程度的拆半可靠性模型。输出结果中给出 Guttman 和 Spearman-Brown 拆半可靠性系数以及拆半后两部分的 Cornbach α 系数，考察两部分间的相关性
Guttman	Guttman 模型法。计算最低下限的真实可靠性法，输出结果中产生 6 个可靠性系数，从 lambda1~lambda6
平行	平行模型法。计算各评估项目变异数同质时的最大概率（maximum-likelihood）可靠性，该模型假设所有项目具有相等的方差和相等的方差误差
严格平行	严格平行模型法。该模型是假设评估项目具有相等均值的平行模型法。输出结果中包含模型的拟合优度检验、各评估项目的方差估计值、项内相关系数、可靠性的无偏估计等统计量

● “统计量”选项

在图 13.2 所示对话框中，单击“统计量”按钮，打开“可靠性分析：统计量”对话框，如图 13.3 所示。该对话框用于选择描述可靠性和评估项目的统计量。

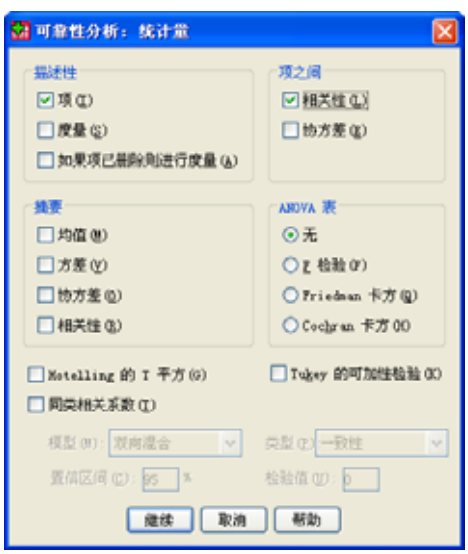


图 13.3 “可靠性分析：统计量”对话框

在图 13.3 所示的对话框中，“ 描述性 ”选项组中给出根据个案产生的尺度或评估项目的描述性统计量，如表 13.2 所示。

表 13.2　描述选项及其含义

选　项	含　义
项	评估项目。选择该项，表示输出各评估项目的基本描述统计量，包括项内均值和标准差等
度量	选择该项，表示输出各评估项目之和的基本描述统计量，包括均值、方差等
如果项已删除则进行度量	选择该项，表示输出剔除某评估项目后的基本统计量，以便对各评估项目逐个评估

“ 摘要 ”选项组中给出根据所有评估项目分布计算的基本描述统计量，如表 13.3 所示。

表 13.3　摘要选项及其含义

选　项	含　义
均值	评估项目均值。输出若干个评估项目平均分的基本描述统计量，包括最大、最小和评估项目均值的平均分，以及评估项目平均分的极差和方差，最大评估项目平均分和最小项目平均分比率
方差	评估项目方差。输出若干个评估项目方差的基本描述统计量，包括最大、最小和评估项目均值的方差，以及评估项目方差的极差和方差，最大评估项目方差和最小评估项目方差的比率
协方差	评估项目的内在协方差。输出若干个评估项目协方差矩阵的基本描述统计量，包括最大、最小和评估项目的内在协方差，以及评估项目内在协方差的极差和方差，最大评估项目内在协方差和最小评估项目内在协方差的比率
相关性	评估项目的内在相关系数。输出若干个评估项目相关系数矩阵的基本描述统计量，包括最大、最小和评估项目的内在相关系数，以及评估项目内在相关系数的极差和方差，最大评估项目内在相关系数和最小评估项目内在相关系数的比率

“ 项之间 ”选项组给出输出统计量的选项，如表 13.4 所示。

表 13.4　项之间选项及其含义

选　项	含　义
相关性	项内相关系数矩阵，表示输出相关系数矩阵的基本描述统计量
协方差	项内协方差矩阵，表示输出协方差矩阵的基本描述统计量

“ ANOVA 表 ”选项组中给出了用于检验同一评估项目在各评估项目上的得分是否具有一致性的方法，有 4 个选项，如表 13.5 所示。

表 13.5　ANOVA 表选项及其含义

选　项	含　义
无	不做检验
F 检验	F 检验。表示重复测量的方差分析，适合于数据为定距型且服从正态分布的情况
Friedman 卡方	Friedman 卡方分析。表示进行非参数检验中的对配对样本 Friedman 检验，适合数据为非正态分布或定序型数据，计算 Friedman 值和 Kendall 一致性系数，在方差分析表中，利用卡方检验代替 F 检验

续表

选　　项	含　　义
Cochran 卡方	计算 Cochran ’s Q 检验值。表示进行非参数检验只能对配对样本进行 Cochran 检验，适合数据为二值型，在方差分析表中，利用 Q 统计量代替 F 统计量

在该对话框中，还给出表 13.6 所示的对评估项目检验和计算的选项。

表 13.6　统计对话框其他选项及其含义

选　　项	含　　义
Hotelling 的 T 平方	检验所有评估项目的均值是否相等的多变量检验
Tukey 的可加性检验	检验评估项目中是否存在倍增交互作用
同类相关系数	进行一致性测度或个案数值的一致性检验。计算组内相关系数需计算方法的相关类型，选择该项时，表 13.7 所示选项变为可用

表 13.7　同类相关系数选项及其含义

选　　项	含　　义
模型	给出用于计算组内相关系数的方法，这里给出 3 种模型，Two-Way Mixed（两方向固定模型，默认选择，当个案效应和评估项目效应均为固定时选择此项）Two-Way Random（两方向随机模型，当个案效应和评估项目效应均随机时选择此项）和 One-Way Random（单方向随机模型，当个案效应为随机时选择此项）
类型	用于选择指标类型，可选类型有两个：ConsistencyAbsolute（一致性，默认选项）Agreement（绝对一致）
置信区间	用于指定置信区间的水平，默认值为 95%
检验值	用于指定假设检验过程的检验量，默认值为 0，可输入 0~1 之间的数值，用于类间相关系数的比较

【示例 13】节目是否继续开办的调查分析

考虑电视台对某一栏目收视情况进行调查的问题。节目策划部门决定下个季度某个节目是否继续开办，是否开办的关键主要是由下个季度该节目的收视率所决定的。该部门设计了一个调查表，希望能通过该调查表中的统计数据得到关于下个季度仍收看该节目的可靠原因。数据如表 13.8 所示。

表 13.8　某节目收视率调查表

any	bored	critics	peers	writers	director	case
0	0	0	0	1	1	1
1	1	0	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
0	0	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

续表

any	bored	critics	peers	writers	director	case
0	1	1	1	1	1	1
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1
0	0	0	0	0	0	0
…	…	…	…	…	…	…

由表 13.8 可知，调查表中包括 7 个项目，问题为：如果满足给定的项目，下个季度你是否观看该节目。给出的 7 个项目为：没有其他任何原因、没有其他流行节目、外界评论仍然很好、有人仍观看该节目、仍是原班编剧、仍是原班导演和仍是原班演员。

要求被访者对每一项目选择“是”或“不是”，数据文件中，“是”记为 1，“不是”记为 0，共收到 906 份有效文件，根据问卷数据和可靠性分析方法对该问卷调查表设计进行可靠性分析。

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开可靠性分析对话框并选入变量。

依次单击菜单“分析→度量→可靠性分析”命令，打开“可靠性分析”对话框。将所有项目选入“项目”列表框，作为评估项目。

其他各项选择默认选项。

03 设置。

单击“统计量”按钮，打开“可靠性分析：统计量”对话框。在“描述性”选项组中选择“项”选项；在“项之间”选项组中选择“相关性”选项。

其他各项选择默认选项。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
RELIABILITY
    /VARIABLES=any bored critics cast peers writers director
    /FORMAT=NOLABELS
    /SCALE(ALPHA)=ALL/MODEL=ALPHA
    /STATISTICS=DESCRIPTIVE CORR
    /ICC=MODEL(MIXED) TYPE(CONSISTENCY) CIN=95 TESTVAL=0 .
```

设置完毕后，单击“确定”按钮，执行可靠性分析操作。

其结果分析如下。

- 个案摘要表

表 13.9　个案摘要表

	N	%
Cases Valid	906	100.0
Excluded[a]	0	.0
Total	906	100.0

a．Listwise deletion based on all variables in the procedure.

表 13.9 显示可靠性分析过程中参与分析的个案数量和缺失值数量。此例共有 906 个个案参与可靠性分析，不含缺失值。

- 评估项目的基本描述统计量

表 13.10　评估项目的基本描述统计量表

Item Statistics

	Mean	Std.Deviation	N
没有其他任何原因	.49	.500	906
没有其他流行节目	.50	.500	906
外界评论仍然很好	.50	.500	906
仍是原班演员	.89	.315	906
有人仍观看该节目	.53	.499	906
仍是原班编剧	.81	.389	906
仍是原理导演	.83	.378	906

表 13.10 是可靠性分析的评估项目基本描述统计量。表中给出了所有项目的均值（Mean）、标准差（Std.Deviation）以及参与分析的个案数（N）。可以看出，表中前 4 个项目的均值在 0.50 左右，即将来愿意继续观看节目的人数占总被访者人数的 50%左右；表中第 3 个项目的均值在 0.80~0.90 之间，从表中项目内容知道，后 3 个项目表示如果项目内容变化不大时，被访者愿意继续观看节目的人数均值。

- 评估项目的相关系数矩阵

表 13.11　评估项目的相关系数矩阵

Inter-Item Correlation Matrix

	没有其他任何原因	没有其他流行节目	外界评论仍然很好	仍是原班演员	有人仍观看该节目	仍是原班编剧	仍是原理导演
没有其他任何原因	1.000	.815	.813	.303	.782	.408	.421
没有其他流行节目	.815	1.000	.826	.307	.807	.422	.423
外界评论仍然很好	.813	.8.26	1.000	.336	.804	.458	.453
仍是原班演员	.303	.307	.336	1.000	.340	.625	.600
有人仍观看该节目	.782	.807	.804	.340	1.000	.443	.460

续表

	没有其他任何原因	没有其他流行节目	外界评论仍然很好	仍是原班演员	有人仍观看该节目	仍是原班编剧	仍是原理导演
仍是原班编剧	.408	.422	.458	.625	.443	1.000	.632
仍是原理导演	.421	.423	.453	.600	.460	.632	1.000

表 13.11 是评估项目的相关系数矩阵。可以看出，前 4 个项目之间的相关系数很大，且是正相关。由此可以得出，如果在其中一个项目条件下愿意将来观看节目的被访者，很有可能在其他 3 个项目中将来也会愿意观看节目。后 3 个项目和其他项目之间的相关系数也为正值。

人们愿意观看某一节目的原因有很多。例如由于生活的无趣等原因使得很多人有可能在下一年继续观看节目；当然，要保证演员的演技、剧本的创作和导演的水平保持在一个时尚和流行的水平。另一方面，现在不观看该节目的人有可能在将来也不会观看该节目。

从上面简单的相关系数矩阵中可以看出很多信息，但对于编制的项目表的可靠性，即可靠性的定量化描述，从相关系数矩阵表中还不能得到。

- 可靠性分析的 Cronbach α 系数

表 13.11 可靠性分析的 Cronbach α 系数

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based On Standardized Items	N of Items
.898	.894	7

Cronbach α 系数是度量可靠性的一种重要方法，关于该系数的含义在前面的章节中已经给予了解释。该例中的 Cronbach α 系数是由表 13.11 的可靠性统计分析表中给出。表中不但给出了 Cronbach α 系数，还给出了评估项目的标准化 Cronbach α 系数；由于可靠性系数为 0.894，大于 0.800，因此总体上该调查评估表编制的内在可靠性是比较理想的。

13.3 可靠性分析的其他问题

可靠性分析除了可以选择 Cronbach α 模型以外，还有拆半可靠性系数模型、Guttman 模型、平行模型和严格平行模型等方法，下面我们利用其他方法分析该可靠性反系问题，通过比较分析可以进一步加深对评估项目的理解。

13.3.1 拆半可靠性系数模型

为了得到一个准确的项目评估可靠性估计值，一个可选择的方法是利用重测可靠性

法，即在一定的时间间隔下，对同一组被访者进行两次相同的问卷调查，分析两次结果之间的相关系数。然而，在实践中这种方法的实现往往受一定的条件限制。为此，我们可以将评估项目拆分为两组评估项目，进行相关性分析和 Guttman 拆半可靠性系数的计算，这就是可靠性分析的另一种重要方法：拆半可靠性分析方法。

可靠性分析操作中，在图 13.2 所示对话框“ 模型 ”下拉菜单中选择“ 半分 ”选项，如图 13.4 所示。单击“ 统计量 ”按钮，打开“ 可靠性分析：统计量 ”对话框，在“ 描述性 ”选项组中选择“ 度量 ”，在“ 项之间 ”选项组中选择“ 相关性 ”，如图 13.5 所示。其他各项设置选择默认选项。

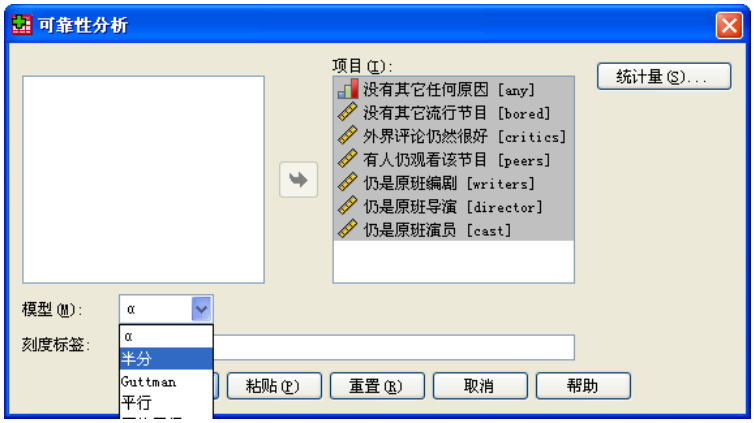


图 13.4　选择“ 半分模型 ”的可靠性分析对话框

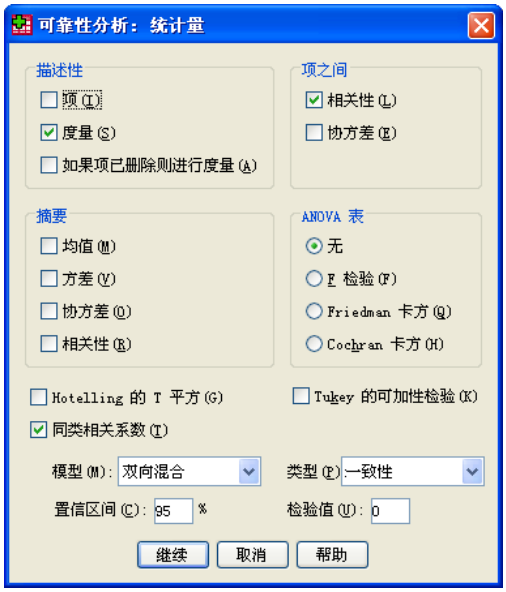


图 13.5　选择“ 半分模型 ”的“ 可靠性分析：统计量 ”对话框

单击图 13.4 对话框中的“粘贴”按钮，得到 Syntax 命令的语句如下：

```
RELIABILITY
    /VARIABLES=any bored critics peers writers director cast
    /FORMAT=NOLABELS
    /SCALE(SPLIT)=ALL/MODEL=SPLIT
    /STATISTICS=SCALE CORR .
```

设置完毕后，单击“确定”按钮，执行拆半可靠性分析操作，如表 13.13 和表 13.14 所示是输出窗口中显示的结果。

表 13.13　拆半可靠性分析的基本描述统计量

Scale Statistics

	Mean	Variance	Std. Deviation	N of Items
Part 1	2.02	3.424	1.850	4[a]
Part 2	2.53	.875	.935	3[b]
Both Parts	4.55	6.040	2.458	7

a．The items are：没有其他任何原因，没有其他流行节目，处界评论仍然很好，有人仍观看该节目。

b．The items are：仍是原班编剧，仍是原班导演，仍是原班演员。

表 13.13 是拆半可靠性分析的基本描述统计量列表。表中给出了拆半后的统计量描述，由于有 7 个评估项目，拆分时按照原顺序，前 4 个作为一组项目，后面的 3 个作为另一组项目，第一组项目的均值为 2.02，方差为 3.424，标准差为 1.850，第二组项目的均值为 2.53，方差为 0.875，标准差为 0.935。

表 13.14 为拆半可靠性分析的可靠性系数。Correlation Between Forms 是两部分量表总分的相关系数，为 0.503，相关程度不算高。而 Spearman-Brown Equal length 系数为 0.669，是利用两部分的标准化 α 系数得到的。Guttman 拆半系数是利用两组项目的 Crobach α x 系数计算得到的，为 0.577。

表 13.14　拆半可靠性分析的可靠性系数

Reliability Statistics

Cronbach's	Part 1	Value	.944
N of Items			4[a]
	Part 2	Value	.826
		N of Items	3[b]
	Total N of Items		7
Correlation Between Forms			.503
Spearman-Brown	Equal Length		.669
Coefficient	Unequal Length		.673
Guttman Split-Half Coefficient			.577

a．The items are：没有其他任何原因，没有其他流行节目，处界评论仍然很好，有人仍观看该节目。

b．The items are：仍是原班编剧，仍是原班导演，仍是原班演员。

注意：项目不同的拆半会产生不同的可靠性系数的估计值。在拆半过程中，如果使具有高相关性的项目分在一组，这时拆半系数将接近于其最小值。当高相关性的项目成对拆分时，拆半系数将达到其最大值。

13.3.2　Guttman 模型

可靠性分析操作中，在图 13.2 所示对话框中选择“ Guttman ”选项作为可靠性分析的模型，其他选项设置选择默认选项。单击“ 粘贴 ”按钮，得到 Syntax 命令，语句如下：

```
RELIABILITY
    /VARIABLES=any bored critics peers writers director cast
    /FORMAT=NOLABELS
    /SCALE(GUTTMAN)=ALL/MODEL=GUTTMAN
    /STATISTICS=SCALE CORR .
```

设置完毕后，单击“ 确定 ”按钮，执行 Guttman 方法可靠性分析操作，表 13.15 和表 13.16 是输出窗口总显示的结果。

表 13.15　Guttman 模型的可靠性分析基本描述统计量

Reliability Statistics

Lambda	1	.769
	2	.915
	3	.898
	4	.577
	5	.894
	6	.927
N of Items		7

The covariance matrix is calculated and used in the analysis.

表 13.15 给出了 6 种不同的 Guttman 可靠性系数。从表中可以看出 Lambda3 要好于 Lambda1，这里的 Lambda3 等价于 Cronbach α 系数。虽然 Lambda2 好于 Lambda1 和 Lambda3，但由于其计算较为复杂，所以在应用中受到一定的限制。

表 13.16　Guttman 模型的基本描述统计量

Scale Statistics

	Mean	Variance	Std.Deviation	N of Items
Part　1	2.02	3.424	1.850	4[a]
Part　2	2.53	.875	.935	3[b]
Both Parts	4.55	6.040	2.458	7

a．The items are：没有其他任何原因，没有其他流行节目，外界评论仍然很好，有人仍观看该节目。

b．The items are：仍是原班编剧，仍是原班导演，仍是原班演员。

表 13.16 给出了计算第一部分和第二部分的基本描述统计量。

13.3.3 平行模型

可靠性分析操作中，在图 13.2 所示对话框中选择“ 平行 ”选项作为可靠性分析的模型。在“ 可靠性分析：统计量 ”对话框“ 描述 ”选项组中选择“ 度量 ”，在“ 项之间 ”选项组中选择“ 相关性 ”，如图 13.5 所示。其他各项设置选择默认选项。

单击“ 粘贴 ”按钮，得到 Syntax 命令，语句如下：

```
RELIABILITY
    /VARIABLES=any bored critics peers writers director cast
    /FORMAT=NOLABELS
    /SCALE(PARALLEL)=ALL/MODEL=PARALLEL
    /STATISTICS=SCALE CORR .
```

设置完毕后，单击“ 确定 ”按钮，执行平行（Parallel）模型可靠性分析操作，表 13.17、表 13.18 和表 13.19 所示的是输出窗口总显示的结果。

表 13.17　平行模型的优度拟合检验

Test for Model Goodness of Fit

Chi-Square	Value	1968.281
	df	26
	Sig	.000
Log of Determinant	Unconstrained Matrix	-16.885
of	Constrained Matrix	-14.704

Under the parallel model assumption

平行（严格平行）模型可以检验均值和方差是否相等。严格平行模型假设真实的项目得分应该具有相等的均值和方差，而平行模型假设相等的方差。表 13.17 给出的就是平行模型的检验结果。从表中可以看出，计算检验统计量的概率 ρ 值为 0.000，远小于 0.05，故拒绝方差相等的假设。

表 13.18　平行模型的可靠性分析结果

Reliability Statistics

Common Variance	.199
True Variance	.111
Error Variance	.088
Common Inter-Item Correlation	.556
Reliability of Scale	.898
Reliability of Scale	.898
(Unbiased)	

The covariance matrix is calculated and used in the analysis.

表 13.18 给出的是平行模型的可靠性分析结果。注意到平行模型的可靠性估计值为

0.898，等价于 Crobachα系数；但由于平行模型假设被拒绝，故严格平行模型自然也被拒绝。实际上，大多数可靠性分析的平行模型的等方差性都不满足，也就是说平行模型假设被拒绝，但该模型仍是有意义的，因为模型结果中提供了其他模型分析中没有的方差估计。

表 13.19　平行模型的基本描述性统计量

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
4.55	6.040	2.458	7

修改：表 13.19 给出了平行模型的基本描述性统计量。从表中可以看出，平行模型可靠性分析的均值为 4.55，方差为 6.040，标准差为 2.458。

13.4　本章小结

本章介绍了可靠性分析的基本方法和 SPSS 操作步骤。重点理解可靠性分析的基本思想和应用范围，掌握可靠性分析的基本原理和基本概念。体会可靠性分析的几种方法，包括常用的 Crobachα系数和拆半可靠性、Guttman 和平行模型可靠性分析方法，要求读者能够理解可靠性分析输出结果的含义。

第14章 时间序列分析

时间序列分析是多元统计分析的一项重要内容。时间序列是指按时间顺序排列的随机变量的一组实测值。分析时间序列图，可以从运动的角度认识事物的本质，如几个时间序列之间的差别、一个较长时间序列的周期性，或对未来的情况进行预测。

在 SPSS 中进行时间序列分析，可以利用“分析→预测”命令实现，菜单内容如图 14.1 所示。

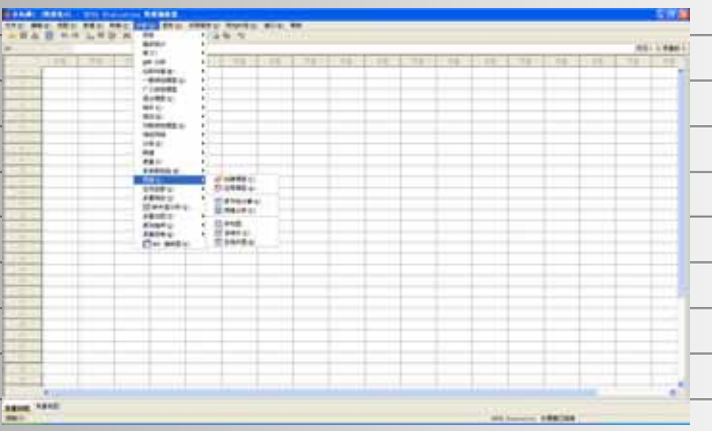


图 14.1　SPSS 时间序列分析模块

本章学习要点：

- 时间序列分析概述；
- 时间序列数据的预处理；
- 指数平滑法；
- 自回归模型；
- ARIMA 模型；
- 周期性分解。

14.1　时间序列分析概述

本节主要讲述时间序列分析的基本内容，包括时间序列数据及其分析方法，时间序列分析的模型、公式和记号等。

14.1.1　时间序列数据及其分析方法

所谓时间序列，是指一个依时间顺序组成的观察数据集合。很多数据以时间序列形式呈现，如货运码头的逐月吞吐量，公路交通事故次数周度报告，城市空气污染物的日均值序列，医院每日门诊接诊人数序列，城市电网每日输电量，地区工业总产值的年度数据序列，逐年人口统计资料等。时间序列区别于普通资料的本质特征是相邻观测值之间的依赖性或称相关性，这种特征使得时间序列资料的统计分析方法区别于一般数据的统计分析方法。事实上，有关时间序列分析的特殊技巧，几乎都是基于对自相关性处理的技巧。

时间序列分析按分析目的不同，可以划分为时域分析和频域分析两个类别，前者将序列的观察值视为历史值的函数，重点分析事物随时间发展变迁的趋势，常用于人口、经济、气象等研究领域；后者将序列看成不同频率的正弦或余弦波叠加的结果，重点分析其频率特征，常用于电力、工程等方面。本章重点介绍时间序列和时域分析方法。

移动平均法、指数平滑法是早期时间序列分析的主流方法。在 20 世纪 70 年代后期，由于 Box 和 Jenkins 的工作及电子计算机的逐渐普及，*ARIMA*（求和自回归滑动平均模型）被大量用于时间序列资料的分析，一般提到的时间序列模型，都是指 *ARIMA* 模型或它的某种表述形式。

预测是时间序列分析的重要内容，几乎所有的时域分析方法，首先都用于预测。主流时间序列分析方法对数据资料要求严格，不允许有缺失值，所以，缺失值填补也是时间序列分析的内容之一，且缺失值填补也是基于预测的。

14.1.2　时间序列分析的模型、公式和记号

1．随机序列、自协方差函数、自相关函数和平稳随机序列的定义。

设 $\{X_t : X_1, X_2, ...\}$ 为随机序列，其二阶原点距大小有限，$EX_k^2 < \infty$（E 表示数学期望，下同），则均值函数 μ_t、自协方差函数 γ_{ts}、自相关函数 ρ_{ts} 有如下定义：

$$\mu_t = \int_{-\infty}^{+\infty} xp(t)dx \tag{14.1}$$

$$\gamma_{ts} = E(X_t - \mu_s)(X_s - \mu_s) \tag{14.2}$$

特定地，当 $t = s$，$\gamma_{ts} = \gamma_{tt} = \mathrm{var}\, X_t$ 时，称 $\mathrm{var}\, X_t$ 为 $\{X_t\}$ 的方差函数。

$$\rho_{ts}=\frac{\gamma_{ts}}{\sqrt{\gamma_{tt}\gamma_{ss}}} \tag{14.3}$$

如果上述随机序列 $\{X_t\}$ 满足：

- 对任意整数 t，$EX_t=\mu$，μ 为常数；
- 对任意整数 t，s，$\gamma_{ts}=\gamma_{t-s}$，即 γ_{ts} 仅与 $t-s$ 有关，而与个别时刻 t，s 无关，则称序列 $\{X_t\}$ 为宽平稳序列，简称为平稳序列。

特别地，当 $EX_t=0$，$\gamma_\tau=EX_tX_s=\begin{cases}\delta^2,t=s\\0,t\neq s\end{cases}$ 时，称序列 $\{X_t\}$ 为白噪声。

2．平稳时间序列——*ARIMA* 过程。

设 $\{x_t\}$ 为零均值平稳序列，$\{a_t\}$ 为白噪声，$Ex_ta_s=0(t<s)$，满足：

$$x_t-\varphi_1x_{t-1}-\varphi_2x_{t-2}-\ldots-\varphi_px_{t-p}=a_t-\phi_1a_{t-1}-\phi_2a_{t-2}-\ldots-\phi_qa_{t-q} \tag{14.4}$$

则 $\{x_t\}$ 为 p 阶自回归——q 阶滑动平均过程，简记为 $ARMA(p,q)$。$\{x_t\}$ 称为 $ARMA(p,q)$ 序列，非负整数 p、q 分别称为自回归阶数和滑动平均阶数，参数 $\varphi_1,\varphi_2,\ldots,\varphi_p$ 称为自回归系数，$\theta_1,\theta_2,\ldots,\theta_p$ 称为滑动平均数。

当 $p=0$ 时，则 $ARIMA(0,q)$ 模型

$$x_t=a_t-\theta_1a_{t-1}-\theta_2a_{t-2}-\ldots-\theta_qa_{t-q} \tag{14.5}$$

称为 q 阶滑动平均模型，记为 $ARI(p)$。

引入后移算子 B，令 $B^kx_t=x_{t-k}$，$B^ka_t=a_{t-k}$，$B^kc=c$（c 为常数），并令

$$\varphi(B)=1-\varphi_1B-\varphi_2B^2-\ldots-\varphi_pB^p$$

$$\theta(B)=1-\theta_1B-\theta_2B^2-\ldots-\theta_pB^p$$

则 $ARIMA(p,q)$ 模型简记为：

$$\varphi(B)x_t=\theta(B)a_t \text{ 或 } x_t=\varphi^{-1}(B)\theta(B)a_t \tag{14.6}$$

若 $\varphi(B)$ 的特征方程

$$1-\varphi_1B-\varphi_2B-\ldots-\varphi_pB^p=0 \tag{14.7}$$

p 个根都在单位圆外，则模型是平稳的。

若 $\theta(B)$ 的特征方程

$$1-\theta_1B-\theta_2B-\ldots-\theta_qB^q=0 \tag{14.8}$$

q 个根都在单位圆外，则模型是可逆的。

3．非平稳时间序列—— *ARIMA* 过程。

定义差分算子 ∇ 为

$$\nabla z_t=z_t-z_{t-1} \tag{14.9}$$

若差分算子 ∇ 和后移算子 B 有以下关系：

$$\nabla=1-B,\nabla^2=(1-B)^2,\nabla^d=(1-B)^d \tag{14.10}$$

称 d 为差分的等价。

设 $\{z_t\}$ 为非平稳序列，$\{x_t\}$ 为 $ARIMA(p,q)$ 序列，存在正整数 d，使得

$$x_t = \nabla^d z_t, t > d \tag{14.11}$$

则有：

$$\varphi(B)(1-B)^d z_t = \theta(B)a_t \tag{14.12}$$

称此模型为求和自回归滑动平均模型，记为 $ARIMA(p,d,q)$。

4．*ARIMA* 模型的识别、参数估计和诊断。

- 自相关函数和偏自相关函数的定义

设 ρ_k 是 $\{x_t\}$ 的自相关函数，则

$$\rho_k = \frac{E(x_t - Ex_t)(x_{t-k} - Ex_{t-k})}{\sqrt{E(x_t - Ex_t)^2 (x_{t-k} - Ex_{t-k})^2}} \tag{14.13}$$

令 $\hat{x}_t \triangleq \hat{E}(x_t \mid x_{t-1}, x_{t-2}, ..., x_{t-k+1})$，$\hat{x}^*_{t-k} \triangleq \hat{E}(x_t \mid x_{t-1}, x_{t-2}, ..., x_{t-k+1})$，记 $\tilde{x}_t = x_t - \hat{x}_t$，$\overline{x}^*_{t-k} = x_{t-k} - \hat{x}^*_{t-k}$。

设 ϕ_{kk} 是 $\{x_t\}$ 的偏自相关函数，则

$$\phi_{kk} = \frac{E\tilde{x}_t \tilde{x}^*_{t-k}}{\sqrt{E\tilde{x}_t{}^2 \tilde{x}^*_{t-k}{}^2}} \tag{14.14}$$

样本 ρ_k 和 ϕ_{kk} 的估计是由 Yule-Walker 方程递推解出的。

- *ARIMA* 模型的识别

根据 Box-Jenkins 提出的方法，用样本的自相关函数和偏自相关函数的截尾初步识别 *ARIMA* 模型的阶数。若平稳序列 $\{x_t\}$ 的 ρ_k 呈 q 步截尾，而 ϕ_{kk} 拖尾，则识别 $\{x_t\}$ 为 $MA(q)$ 序列；若 ρ_k 拖尾，而 ϕ_{kk} 呈 p 步截尾，则识别 $\{x_t\}$ 为 $ARI(p)$ 序列；若 ρ_k、ϕ_{kk} 均拖尾，则判断 $\{x_t\}$ 为 $ARIMA(p,q)$ 序列。首先可以经验性地给出 p、q 的初步识别，然后通过模型诊断反复识别，找到最优的 p、q 组合来确定。

- 参数估计

参数估计即为 *ARIMA* 模型的条件似然函数和条件最小二乘估计。*ARIMA* 模型可表示为以下差分方程形式：

$$a_t = w_t - \phi_1 w_{t-1} - ... - \phi_p w_{t-p} + \varphi_1 w_{t-1} + ... + \varphi_q w_{t-q} \tag{14.15}$$

其中，$w_t = \nabla^d z_t, t = 1,2,...n$。

在给定 w 的 p 个初始值 w^* 和 a 的 q 个初始值 a^* 的条件下，对于任意给定的参数 (ϕ,θ) 和初始值 (w^*,a^*)。假设 a 为正态分布，其概率密度为：

$$p(a_1,a_2,...,a_n) \propto \delta_a^{-n} e^{[-(\sum_{t=1}^{n} \frac{a_t^2}{2\delta_a^2})]} \tag{14.16}$$

则与参数 $(\varphi,\theta,\delta_a)$ 相联系的对数条件似然函数是：

$$l_*(\varphi,\theta,\delta_a^2) = -n\ln(\delta_a^2) - \frac{S*(\varphi,\theta)}{2\delta_a^2} \tag{14.17}$$

其中：

$$S^*(\varphi,\theta)=\sum_{t=1}^{n}a_t^2(\varphi,\theta\mid w^*,a^*,w) \tag{14.18}$$

$S^*(\varphi,\theta)$ 称为条件平方和函数，使其最小化，可得参数的条件最小二乘估计。

● 模型的诊断：残差的自相关检验

对模型的残差序列计算其 m 个自相关函数估计值 $\hat{\rho}_k$，构造统计量

$$Q=n\sum_{k=1}^{m}\hat{\rho}_k \tag{14.19}$$

它服从 $v=m-p-q$ 的 χ^2 分布。

5．带有 *ARIMA* 误差的回归模型—— *ARIMA* 模型。

如果一个回归方程的误差是一个 *ARIMA* 过程，则称此模型为带有 *ARIMA* 误差的回归模型。本模型的数学表示为：

$$y_t=X_{t-b}\beta+N_t \tag{14.20}$$

X 为输入序列或解释变量序列，b 为滞后参数，如果 b 为零，则输入序列的效应即反映在 y 序列上；如果 b 大于零，则表示输入序列的效应经过 b 个滞后才在 y 序列上体现。β 为回归系数或回归系数变量，N_t 为系统噪声，如果 N_t 为 *ARIMA* 噪声，即 $N_t=\varphi^{-1}(B)\theta(B)a_t$，其中 a_t 为白噪声，此模型则称为附加 *ARIMA* 噪声的回归模型，简称为 *ARIMA* 模型。

实际上，带有 *ARIMA* 误差的回归模型是一种简单的传递函数模型（Transfer Function Model）。当输入序列对响应序列的作用为有一定时滞的累积效应时，使用传递函数更为方便。对于动态系统，用线性近似来刻画输出 y_t 和输入 x_t 的关系时，可以用一个线性滤波器来表示。

$$y_t=v_0x_t+v_1x_{t-1}+\ldots=(v_0+v_1B+\ldots)x_t=v(B)x_t \tag{14.21}$$

在 14.21 式中，某时刻 t 的输出表示成时刻 $t,t-1,\ldots$ 输入的线性组合，算子 $v(B)$ 称作传递函数。

传递函数的因子表示可以得到参数简化的传递函数形式，即：

$$v(B)=\frac{w_0+w_1B+\ldots+w_sB^s}{1-\delta_1B-\ldots-\delta_rB^r}=\frac{w(B)}{\delta(B)} \tag{14.22}$$

这样就将 $v(B)$ 的无穷多个参数简化成为 $r+s+1$ 个参数。

SPSS 的 *ARIMA* 过程没有提供估计形如式（14.22）的传递函数的功能，对于有累积响应的情况，仍然可以设法拟合简单的传递函数模型。例如，如果有理由认为某输入序列的效应既有瞬时响应又在 3 个时滞内有累积响应（如提高利用率，导致企业投资在当年以及随后的 3 年总萎缩），那么可以利用原始输入序列产生 3 个滞后序列，用这

4 个序列和响应序列建立 *ARIMA* 模型。和传递函数模型相比，这种建模方式简便、容易理解，使用 SPSS 完全可以实现；缺点是可能会出现参数冗余，参数假设检验不易有统计学意义。

识别传递函数或输入序列的时滞效应的基本工具是输入和输出序列的互协方差函数和互相关函数，对输入序列的预白噪化处理可以简化识别传递函数的过程。一旦传递函数的形式得到确定，即可通过条件最小二乘法对传递函数和噪声进行拟合。

6．季节模型。

时间序列常呈周期性变化，或称为季节性趋势。用普通的 *ARIMA* 模型处理这种季节性趋势会导致参数过多，模型复杂。季节性乘积模型可以得到参数简约的模型。季节性乘积模型表示为：

$$\phi_p(B)\Phi_p(B^s)\nabla^s\nabla_d^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{14.23}$$

其中，p，d，q 保持原有含义，P，D，Q 分别表示以 s 为间距的自回归、差分和移动平均算子的阶数，s 为季节参数，如果是月度资料，要描述年度特征，则 $s=12$；如果是日志资料，欲描述周度特征，则 $s=7$。季节性乘积模型简记为 $(p,d,q)\times(P,D,Q)_s$。

14.1.3　SPSS 时间序列分析功能

SPSS 时间序列分析的主模块是“分析”中的“时间序列”模块，提供“指数平滑”、“自回归”、“ARIMA 模型”和“季节分析”。

时间变量的定义由“数据”菜单下的“定义日期”完成。在“转换”菜单下，“日期/时间”提供对时间变量的运算功能；“创建时间序列”提供时间序列的有关计算功能，如产生差分分析、移动序列分析、滞后序列和进行序列修匀等处理；“替换缺失值”提供缺失值填补功能。

在“图表”菜单下，“时间序列”子菜单有“自相关”、“互相关”和“谱密度”三个下拉菜单，分别提供（偏）自相关图、相关图和谱密度（周期）图分析功能。另外，“序列”子菜单提供了时间序列数据的专用线图作图功能。

14.2　时间序列数据的预处理

时间序列数据和普通数据不同，它有严格的顺序，并且需要定义时间变量让软件读懂其时间顺序，特别是季节性模型，必须使用 SPSS 软件内部的时间变量。一些时间序列分析方法（如自回归模型）要求数据没有缺失值，通常在时间序列分析前需要对数据填补缺失值。另外，根据时间序列的顺序特定，可以产生移动平均序列、滞后或领先序列，这些都属于时间序列资料的预处理工作。

14.2.1 定义日期变量

定义日期模块（如图 14.2 所示，依次单击菜单“ 数据→定义日期 ”命令可以得到“ 定义日期 ”对话框）可以产生周期性的时间序列日期变量。使用定义日期对话框定义日期变量，需要在数据窗口读入一个按某种时间顺序排列的数据文件，数据文件中的变量名不能与系统默认的时间变量名重名，否则系统建立的日期变量会覆盖同名变量。系统默认的变量名有：年份，年份、季度，年份、月份，年份、季度、月份，日，星期、日，星期、天（5），星期、天（6）小时，日、小时等。

图 14.2 中有两个栏目，“ 第一个个案为 ”列表框定义时间变量的间隔，“ 第一个个案为 ”要求填入相应的起始日期值，当选定时间间隔和起始日期后，系统便自动按选定的时间间隔产生相应的日期变量。“ 当前日期 ”在图 14.2 的左下方，显示已经存在的日期变量。如果数据集从未定义过日期变量，则“ 当前日期 ”为“ 无 ”。

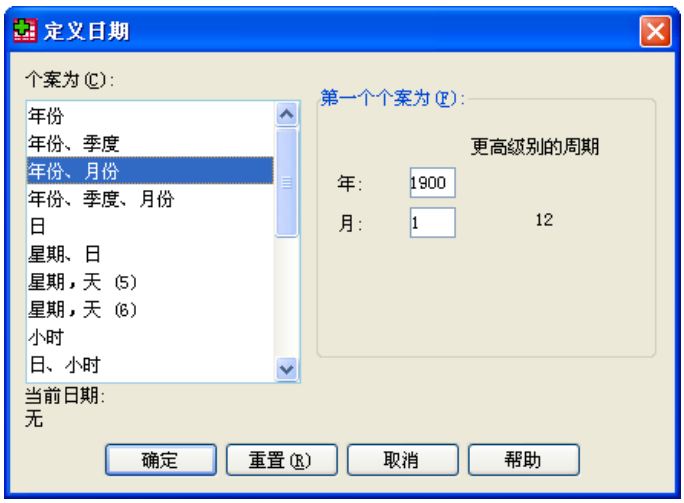


图 14.2 “ 定义日期 ”对话框

定义日期变量的具体操作过程如下。

01 选择定义日期变量的对话框。

依次单击菜单“ 数据→定义日期 ”命令，打开“ 定义日期 ”对话框，如图 14.2 所示。

02 选择日期格式。

在图 14.2 所示的对话框的左侧列表框中选择日期格式。

03 定义数据的起始值。

在图 14.2 所示的对话框中定义数据的起始年份、月份。

定义完毕后，单击“ 确定 ”按钮，在数据集中生成日期变量。

在图 14.2 所示对话框左侧列表框中为观测量的选项，其含义如表 14.1 所示。

表 14.1　观测量选项及其含义

选　项	含　义
年份	时间间隔为年份
年份，季度	时间间隔为季度，以年为周期
年份，月份	时间间隔为月份，以年份为周期
年份，季度，月份	时间间隔为月份，以季度和年份为周期
日	时间间隔为日
星期，日	时间间隔为日，以星期为周期
星期，工作日(5)	时间间隔为工作日（5 天工作日），以星期为周期
星期，工作日(6)	时间间隔为工作日（6 天工作日），以星期为周期
小时	时间间隔为小时
日，小时	时间间隔为小时，以天为周期
日，工时 (8)	时间间隔为小时（8 小时工作制），以天为周期
星期，日，小时	时间间隔为小时，以天、周为周期
星期，工作日，小时	时间间隔为小时，以工作日、天为周期
分钟	时间间隔为分钟
小时，分钟	施加间隔为分钟，以小时为周期
日，小时，分钟	时间间隔为分钟，以小时、天为周期
秒钟	时间间隔为秒
分钟，秒钟	时间间隔为秒，以分钟为周期
小时，分钟，秒钟	时间间隔为秒，以分钟、小时为周期
未注日期	删除已有的时间变量
设定(C)	通过命令语句（编程）产生日期变量

14.2.2　创建时间序列

创建时间序列模块（如图 14.3 所示，依次单击菜单“ 转换→创建时间序列 ”命令可以得到该对话框）的功能为以现有数据为基础，产生新的时间序列，如差分序列、领先序列、滞后序列和移动平均序列等。差分序列主要用于序列的平稳化变换，领先和滞后序列主要用于探讨多个序列间的关系并以此建立模型。移动平均序列主要用于修匀和提供简单的预测功能。

在图 14.3 中，左侧为数据集的变量列表，前 4 个变量为原始变量，后两个变量为通过“ 定义日期 ”模块定义的 SPSS 系统日期变量。变量列表中列出的是变量的标签，最后的变量为“ 月（Month）”，有“ period12 ”的描述，表示周期有 12 个月，即 1 年。图 14.3 的右侧从上到下分为两部分：上面的“ 变量→新名称 ”为新生成的序列的序列（变量）名及其描述，下面的“ 名称和函数 ”为指定新序列变量名和产生新序列的规则。新序列的变量名系统默认为原始变量+下画线+序号。用户也可以自行命名，单击“ 更改 ”按钮完成。“ 函数 ”为一下拉菜单，定义了产生新序列的规则（函数）。“ 顺序 ”和“ 跨度 ”为与有关函数

相关的参数，需要填写。最后一项“ 当前日期 ”指明当前数据的时间周期。

创建时间序列的具体操作过程如下。

01 选择创建时间序列的对话框。

依次单击菜单“ 转换→创建时间序列 ”，可以打开“ 创建时间序列 ”对话框，如图 14.3 所示。

02 设置。

在图 14.3 所示的对话框中选择函数、指定函数参数、选择待处理的原始序列。

定义完毕后，单击“ 确定 ”按钮，在数据集中生成日期变量。

在图 14.3 所示对话框“ 函数 ”下拉菜单中有很多函数可以选择，其选项及含义如表 14.2 所示。



图 14.3 “ 创建时间序列 ”对话框

表 14.2 函数选项及其意义

选　项	含　义
差值	产生差分序列
季节性差分	产生季节性差分序列，需要在“ 次序 ”文本框中输入差分的阶
中心移动平均	产生以当前值为中心的移动平均序列，需要在“ 跨度 ”文本框中输入宽度参数
先前移动平均	产生以当前值之前的数个相邻的值计算的移动平均序列，需要在“ 跨度 ”文本框中输入宽度参数，在序列的开始处会产生和窗宽相等数目的缺失值
运行中位数	类似中心移动平均，只不过此处计算的是相应的中位数
累计求和	计算累积和序列（当前值及所有历史值之和）
滞后	产生滞后序列
提前	产生领先序列
平滑	产生基于混合数据平滑法计算的平滑序列

【示例 14.1】时间序列的预处理。

观察表 14.3 数据中变量“ 年末人口数 ”的 1、2 阶差分序列。

表 14.3　上海市 1978 年～2004 年人口、经济统计资料

年度	年末人口数（万人）	非农业人口数（万人）	人口密度（人/平方公里）	财政收入（亿元）	财政支出（亿元）	生产总值（亿）	人均生产总值（元）	税收（亿元）
1978	1098.28	645.23	1776.00	190.67	26.01	272.81	2498.00	51.51
1979	1132.14	687.38	1830.00	192.75	27.06	286.43	2568.00	53.73
1980	1146.52	702.43	1854.00	198.85	19.18	311.89	2738.00	57.59
1981	1162.84	715.08	1880.00	204.52	19.06	324.76	2813.00	62.21
1982	1180.51	731.31	1908.00	200.69	20.68	337.07	2877.00	65.00
1983	1194.01	745.86	1930.00	204.34	22.39	351.81	2963.00	67.04
1984	1204.78	760.75	1948.00	215.79	30.32	390.85	3259.00	76.90
1985	1216.69	776.37	1967.00	263.86	46.07	466.75	3855.00	102.16
1986	1232.33	802.56	1944.00	257.72	59.08	490.83	4008.00	108.66
1987	1249.51	822.31	1971.00	241.36	53.85	545.46	4396.00	114.00
1988	1262.42	838.93	1991.00	261.69	65.88	648.30	5161.00	126.72
1989	1276.45	855.84	2013.00	297.25	73.31	696.54	5489.00	143.23
1990	1283.35	864.46	2024.00	284.36	75.56	756.45	5910.00	152.37
1991	1287.20	869.88	2030.00	324.66	86.05	893.77	6955.00	161.17
1992	1289.37	875.55	2034.00	340.13	94.99	1114.32	8652.00	182.63
1993	1294.74	893.46	2042.00	439.53	129.26	1511.61	11700.00	255.70
1994	1298.81	910.49	2048.00	615.91	196.92	1971.92	15204.00	179.95
1995	1301.37	921.70	2052.00	702.46	267.89	2462.57	18942.00	226.72
1996	1304.43	932.14	2057.00	873.76	342.66	2902.20	22275.00	271.28
1997	1305.46	943.03	2059.00	1070.95	428.92	3360.21	25750.00	303.64
1998	1306.58	953.65	2061.00	1146.00	480.70	3688.20	28240.00	339.34
1999	1313.12	969.63	2071.00	1390.58	546.38	4034.96	30805.00	365.29
2000	1321.63	986.16	2084.00	1752.70	622.84	4551.15	34547.00	417.00
2001	1327.14	999.07	2093.00	1995.62	726.38	4950.84	37382.00	458.28
2002	1334.23	1018.81	2104.00	2202.25	877.84	5408.76	40646.00	554.70
2003	1341.77	1041.39	2116.00	2828.87	1102.64	6250.81	46718.00	686.64
2004	1352.39	1097.60	2133.00	3325.14	1395.69	7450.27	55307.00	842.74

其在 SPSS 中的具体操作为：打开数据文件，依次单击菜单“ 转换→创建时间序列 ”命令，将图 14.3 中左侧列表框中的变量“ X（年末人口数）”选中，单击 按钮，将其选入到“ 变量→2 新名称 ”列表框。选定“ 函数 ”为差分“ 差值 ”，差分的阶“ 顺序 ”设定为 1 或 2，规定差分序列的变量名（也可以由程序自动生成），然后单击“ 更改 ”按钮，再单击“ 确定 ”按钮。

在 SPSS 数据表中，发现添加了新的变量“ X_1 ”为差分的阶，如图 14.4 所示，这就是原始变量 *X* 的差分序列。

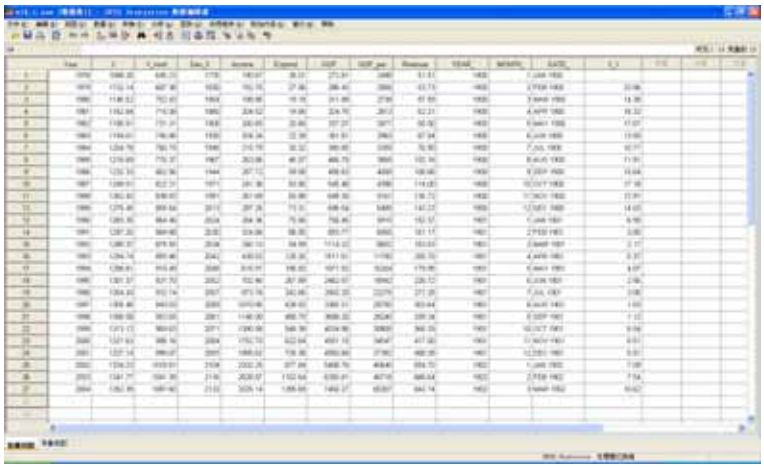


图 14.4　生成新的变量

可以通过做时间序列的序列图（Sequence Chart）来观察原始序列和差分序列的形态，初步认识序列特征。具体操作为：依次单击菜单“ 图形→图表构建程序 ”命令，在打开的对话框中将原始变量或其差分序列选入“ 变量 ”列表框，将时间变量“ 年份（Year）”选入“ 时间坐标轴标签 ”，单击“ 确定 ”按钮，则生成如图 14.5 所示的原始序列及其阶差分序列。

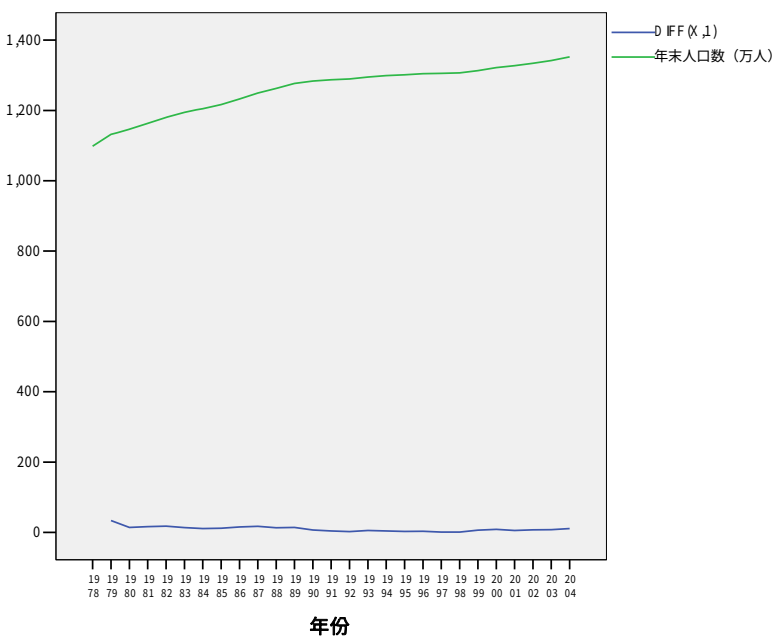


图 14.5　原始序列及其 1 阶差分序列

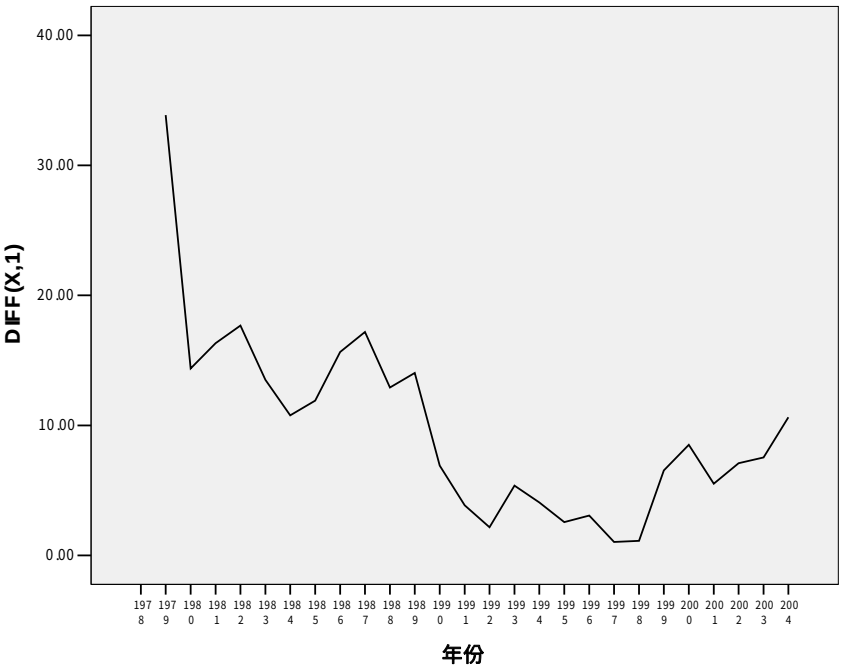


图 14.5　原始序列及其 1 阶差分序列（续）

14.2.3　填补缺失数据

填补缺失数据是时间序列资料分析的重要环节。时间序列分析的参数模型，如 *ARIMA* 模型等，都不允许有缺失值存在，在有缺失值的情况下，系统会用默认的方式填补后分析。SPSS 提供了缺失值填补模块，数据分析者可以选择填补缺失数据的方式。

【示例 14.2】填补缺失数据示例。

表 14.4 为某市连续 60 天日平均气温，*Z* 为原始序列，*Zm* 为模拟有 3 个缺失数据的序列，试对序列 *Zm* 用不同的方法填补并比较结果。

表 14.4　某市连续 60 天平均气温数据

编　　号	*Z*	Z_m
1	22.80	22.80
2	23.90	23.90
3	25.10	25.10
4	26.30	26.30
5	27.90	27.90
6	26.20	26.20
7	26.30	26.30
8	27.00	27.00

续表

编　　号	Z	Z_m
9	23.00	23.00
10	23.40	23.40
11	22.20	22.20
12	23.90	.
13	26.30	26.30
14	28.10	28.10
15	29.50	29.50
16	30.00	30.00
17	30.10	30.10
18	30.30	30.30
19	31.70	31.70
20	27.40	27.40
…	…	…

其在 SPSS 中的具体操作为：打开数据文件，依次单击菜单“转换→替换缺失值”命令，在“替换缺失值”列表框左侧选中变量“zm”，然后选择“方法”下拉菜单中的方法，单击 按钮，将其选入到“新变量”列表框，如图 14.6 所示。



图 14.6　“替换缺失值”对话框

“方法”下拉菜单中各个选项的含义和使用如表 14.5 所示。

表 14.5　方法选项及其含义

方 法 名	含　　义
序列均值	使用全序列均值填补缺失值
临近点的均值	使用以缺失值为中心的移动平均数来填补缺失值，需要在“近点的跨度”文本框中输入参数，实际是半个窗宽

续表

方 法 名	含　义
临近点的中位数	使用以缺失值为中心的中位数来填补缺失值，其他同上
线性插值法	使用线性插值法，即使用半窗宽为 1 的移动平均数插值填补
点处的线性趋势	使用时间变量对原始数据做线性回归，然后根据线性回归方程的预测值填补缺失值

单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
RMV
    /zm_1=SMEAN(zm)
    /zm_2=MEAN(zm 2)
    /zm_3=LINT(zm)
    /zm_4=TREND(zm)
    /zm_5=MEDIAN(zm 2).
```

设置完毕后，单击“ 确定 ”按钮，执行填补缺失值操作，结果如图 14.7 所示。

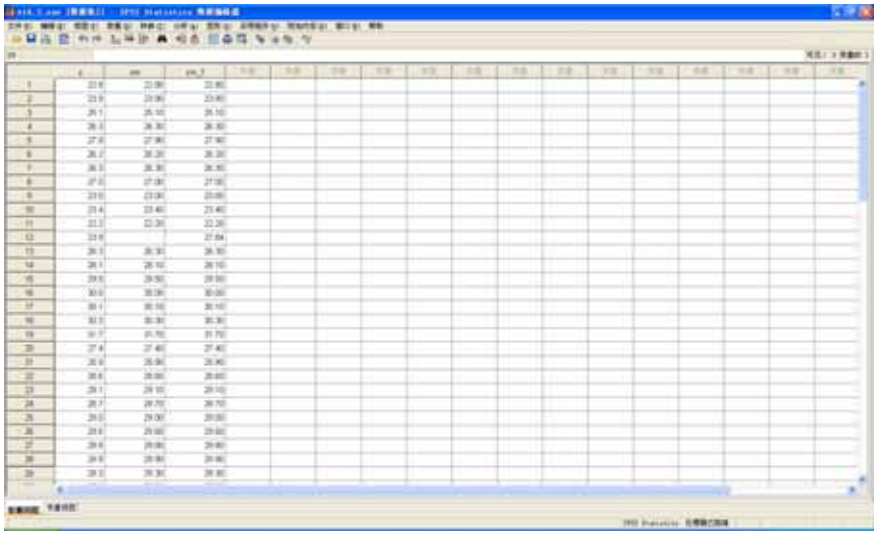


图 14.7　填补后缺失值

从图 14.7 可以看出，填补后的值出现在原始数据的后面。读者可以对其进行对比分析。关于更多“ 替换缺失值 ”模块的操作请读者参阅其他资料，将不作为本章内容重点讲述。本例分别使用了 5 种方法填补缺失值。由于原始序列日平均气温在一段时间内呈线性趋势，所以使用序列均值填补的误差比较大，而其他 4 种方法填补效果较好。

14.3　指数平滑法

指数平滑法的思想来源于对移动平均法预测方法的改进。用当前值和历史值预测未来值时，移动平均法（Point Moving Average）有两个难题：其一是当前值和历史值同等权重

不合理，一般而言，未来值总是和邻近时点的值关系更密切；其二是无法令人信服地确定窗宽，使用 5 日移动平均数还是 15 日这一类难有定论，而且，如果使用 5 日移动平均数，那么 5 日之前的观察值等于赋予权重 0，而 5 日内的观察值均有相等权重 0.2，这也和实际情况相悖。指数平滑法的思想是使用无穷大为宽度，各历史值的权重随时间的推移呈指数衰减，这样就解决了移动平均法的两个难题。指数平滑法可用公式表达如下：

$$\hat{z}_{t+1}=\frac{\sum_{j=0}^{\infty}\theta^{j}z_{t-j}}{\sum_{j=0}^{\infty}\theta^{j}}=(1-\theta)\sum_{j=0}^{\infty}\theta^{j}z_{t-j} \qquad (14.24)$$

其中，$0\leqslant\theta\leqslant1$；$j=0,1,2...$；$t=1,2,...$；$t>j$。

时间序列自身一般有随机波动、长期（线性或非线性）趋势和周期性（稳定性或不稳定性）波动三方面特征，SPSS 软件提供了 3 种指数平滑模型和 1 个自定义模块来处理相应的时间序列，具体如表 14.6 所示。

表 14.6　指数平滑法方法

方　法	说　明
Simple 法	本法为单参数的指数平滑模型，适用于无长期趋势和周期性波动的序列
Holt 法	本法称为双参数线性指数平滑法，适用于有线性趋势而无季节性趋势的时间序列
Winters 法	本法为 3 参数模型，适用于有周期性变化的时间序列数据
Custom 法	本法是 SPSS 提供的一个选项，并非是一种简单的方法，而是提供给用户的自定义方法集合。本法既可以解决以上 3 种方法所解决的问题，又可以处理呈指数趋势或趋势逐渐衰减的序列，还可以针对加法型周期变化和乘法型周期变化分别处理

【示例 14.3】试用不同的指数平滑法对表 14.3 中变量年末人口数进行预测。

图 14.5 中的上图为变量年末人口数在 25 年间的趋势线图，从此图可以看出，年末人口数呈逐年增加趋势，开始增长较快，而后变慢，近似线性趋势，也可以说呈衰减的线性趋势，或者用指数趋势描述更精确。所以，试用 Linear、Exponential 和 Damped 三种方法进行预测并选择最优预测模型和参数。

本例 SPSS 操作的具体步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开指数平滑对话框和选入变量。

依次单击菜单“分析→预测→指数平滑”命令，打开“指数平滑”对话框。将“年末人口数（万人）”选入“类别”列表框，如图 14.8 所示。

03 设置。

在图 14.8 所示对话框“模型”选项组中单击“自定义”按钮，打开“指数平滑：自定义模型”对话框，在该对话框的“趋势分析”选项组中选择“线性”选项，在“季节分析”

选项组中选择“ 无 ”选项，如图 14.9 所示。

在图 14.8 所示对话框中单击“ 参数 ”按钮，打开“ 指数平滑：参数 ”对话框，如图 14.10 所示。在“ 一般（Alpha）”选项组中选择“ 网格搜寻 ”；同样；在“ 趋势（Gamma）”选项组中也选择“ 网格搜寻 ”。

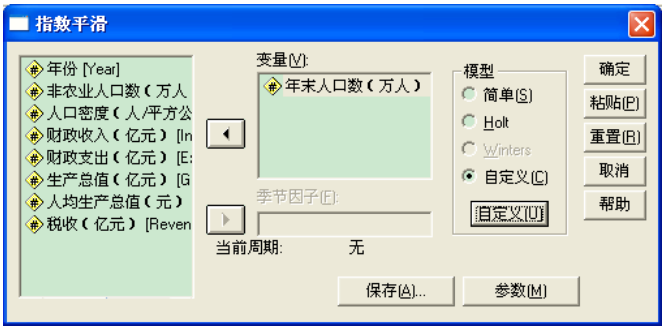


图 14.8 “ 指数平滑 ”对话框

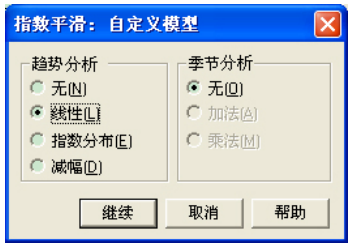


图 14.9 “ 指数平滑：自定义模型 ”对话框

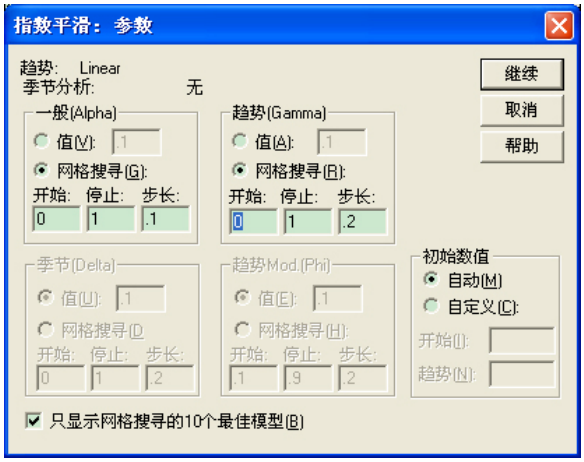


图 14.10 “ 指数平滑：参数 ”对话框

04 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
* Exponential Smoothing.
   TSET PRINT=DEFAULT NEWVAR=ALL .
   PREDICT THRU END.
   EXSMOOTH /VARIABLES=X
   /MODEL=LN
   /ALPHA=GRID(0 1 .1)
   /GAMMA=GRID(0 1 .2)
   /INITIAL=CALCULATE.
```

设置完毕后，单击“ 确定 ”按钮，执行指数平滑操作。

预测序列和预测误差序列会在原始数据后添加，不同参数组合下的均方误差在 Output

窗口给出，并列出最优参数组合。

继续使用“指数分布”和“减幅”选项进行预测，注意：在使用减幅法时，参数ϕ不能为 0 或 1，只能取两者之间。得到其结果分析如表 14.7 所示。

表 14.7　3 种方法预测的结果：最优参数组合表和误差平方和表

（a）线性法

Model Description

Model Name		MOD_3
Series	1	年末人口数（万人）
Holt's Model	Trend	Linear
	Seasonality	None

Applying the model specifications from MOD_3

Smoothing Parameters

Series	Alpha (Level)	Gamma(Trend)	Sums of Squared Errors	df error
X	1.00000	.40000	1006.701	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

（b）指数分布法

Model Description

Model Name		MOD_4
Series	1	年末人口数（万人）
Holt's Model	Trend	Exponential
	Seasonality	None

Applying the model specifications from MOD_4

Smoothing Parameters

Series	Alpha (Level)	Gamma(Trend)	Sums of Squared Errors	df error
X	.70000	1.00000	794.41921	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

（c）减幅法

Model Description

Model Name		MOD_5
Series	1	年末人口数（万人）
Holt's Model	Trend	Damped
	Seasonality	None

Applying the model specifications from MOD_5

Smoothing Parameters

Series	Alpha (Level)	Gamma(Trend)	Phi (Trend Mod.)	Sums of Squared Errors	df error
X	1.00000	1.00000	.10000	4459.250	25

Shown here are the parameters with the smallest Sums of Squared Errors. These parameters are used to forecast.

14.4　自回归模型

在做线性回归分析时，有一个前提条件就是要模型残差相互独立。一些按时间顺序搜集的资料往往存在自相关性，表现为模型的残差间存在自相关现象。这类资料可以使用自回归模型（Autoregression Model）进行分析。

本过程相当于在普通回归方程的右边添加 1 阶自回归算子。自回归模型的数学表达式如下：

$$Y_t = \phi_1 Y_{t-1} + X_t \beta + \varepsilon_t \qquad (14.24)$$

其中，ϕ_1 为自回归系数，Xt 为解释变量序列或自变量序列（可以是 1 个变量，也可以是多个），β 为回归系数向量，ε_t 为白噪声序列。

实际上，单纯的自回归模型并不需要解释变量序列，即形如 $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$ 的模型就是自回归模型的基本形式。但是，SPSS 自回归过程必须要输入自变量，否则不能运行。如果要求不带自变量序列的自回归模型，可以用“ 创建时间序列 ”模块创建一步滞后序列，即 Y_{t-1} 序列，然后把 Y_{t-1} 序列当成自变量即可。

按照时间顺序整理好数据（这点很重要，否则可能出现完全错误的结果），或者定义 SPSS 内部时间变量，然后依次单击菜单“ 分析→预测→自回归 ”命令，打开“ 自回归 ”对话框，如图 14.11 所示。

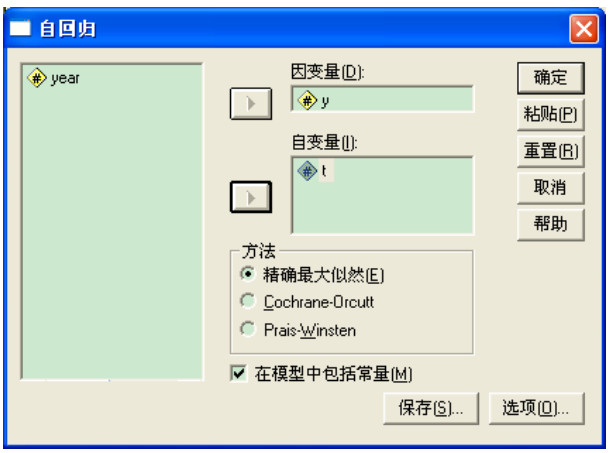


图 14.11　“ 自回归 ”对话框

图 14.11 右侧“ 因变量 ”列表框选入因变量（必须有一个因变量），“ 自变量 ”列表框中选入自变量或解释变量（至少需要 1 个自变量）。“ 方法 ”选项组中提供了 3 种参数估计方法，如表 14.8 所示。

注意：在图 14.11 所示的“ 方法 ”选项组中，无论是 Cochrane-Orcutt 法还是 Prais-Winsten 法都不能处理有缺失值的数据，这两种方法给出的结果和回归过程给出的结果相似，除参

数估计外，有 R^2、方差分析、Dubin-Waterson 统计量等结果；而“精确最大似然估计法”则可给出参数的相关矩阵和协方差矩阵等附加结果。

表 14.8　方法选项及其说明

方 法 名	说　明
精确最大似然	默认选项，本方法允许数据有缺失值
Cochrane-Orcutt	由 Cochrane-Orcutt 于 1949 年提出的基于普通最小二乘法，是针对回归残差自相关现象的处理算法
Prais-Winsten	该方法由 Prais-Winsten 于 1953 年提出，对 Cochrane-Orcutt 算法提出了改进

选定方法以后，依次打开“自回归：保存”和“自回归：选项”对话框，勾选有关选项，如图 14.12 和图 14.13 所示。



图 14.12　“自回归：保存”对话框



图 14.13　“自回归：选项”对话框

在“自回归：保存”对话框中各选项及其含义如表 14.9 所示。

表 14.9　“自回归：保存”对话框选项及其含义

选　项	含　义
创建变量	创建新变量（预测值或拟合值、残差等）的方式。有三个单选按钮，分别是“添加到文件”、“替换现有的”和“不创建”
预测观测量	指定预测记录长度，如果前面选择不创建新变量，则不选择此项。该选项包含两个单选按钮，分别是“预测从估计周期到最后观测量”和“预测经过”。在选择“预测经过”单选按钮时，激活“Obervation”文本框，用户可以输入预测的记录数

“自回归：选项”对话框中各选项及其含义如表 14.10 所示。

表 14.10　“自回归：选项”对话框选项及其含义

选　项	含　义
收敛标准	指定收敛判别标准，一般使用默认即可
显示	规定输出结果，包括“初始和最终参数，以及迭代摘要”、“初始和最终参数，以及详细迭代情况”和“仅仅最终参数”三个单选按钮

【示例 14.4】自回归示例。

表 14.11 为某医院 1981～2003 年恶性肿瘤住院人数资料，其中 y 为住院人数，数据已

按时间顺序排列。初步观察，恶性肿瘤住院人数随时间有线性增长趋势，试做回归分析。

表 14.11　某医院 1981～2003 年恶性肿瘤住院人数数据

年　份	序　号	住院人数 y	年　份	序　号	住院人数 y
1981	1	274	1993	13	888
1982	2	270	1994	14	889
1983	3	253	1995	15	963
1984	4	331	1996	16	1033
1985	5	526	1997	17	938
1986	6	392	1998	18	937
1987	7	365	1999	19	923
1988	8	461	2000	20	1024
1989	9	516	2001	21	1162
1990	10	668	2002	22	1374
1991	11	653	2003	23	1481
1992	12	705			

本例进行 SPSS 操作的具体步骤如下。

01 首先想到的是建立住院人数对时间的回归方程，一般也都是这么开始分析的。不妨先做普通回归分析，看看结果如何。因为年份等间隔增加，为了计算结果简便，将原始数据的时间变量由实际年改成顺序号 1~23。

02 做普通回归分析。以 y 为因变量，以 t 为自变量，计算 Durbin-Watson 统计量，要求将预测值添加到原始数据中。具体操作为：依次单击菜单“ 分析→回归分析→线性 ”命令，打开“ 线性回归 ”对话框，选择 y 为因变量，t 为自变量；然后单击“ 统计量 ”按钮，在“ 残差 ”选项组中选择“ Durbin-Watson ”；再单击“ 保存 ”按钮，在“ 预测值 ”选项组中选择“ 未标准化 ”。分别如图 14.14、图 14.15 和图 14.16 所示。

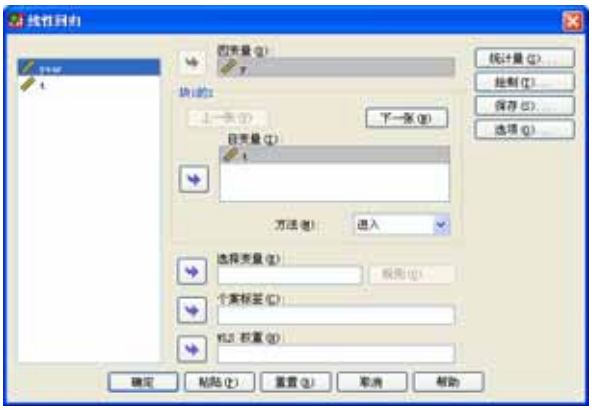


图 14.14 “ 线性回归 ”对话框

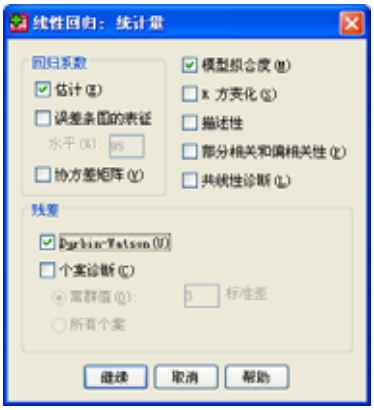


图 14.15 “ 线性回归：统计量 ”对话框

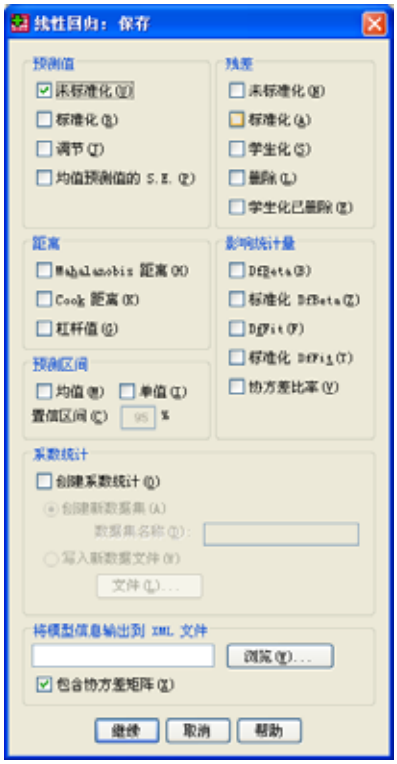


图 14.16 “ 线性回归：保存 ” 对话框

03 单击图 14.14 对话框上 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
REGRESSION
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS R ANOVA
   /CRITERIA=PIN(.05) POUT(.10)
   /NOORIGIN
   /DEPENDENT y
   /METHOD=ENTER t
   /RESIDUALS DURBIN
   /SAVE PRED .
```

04 设置完毕后，单击 “ 确定 ” 按钮，执行线性回归分析操作。得到其结果分析如表 14.12，表 14.13 和图 14.17 所示。

表 14.12　模型摘要

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin- Watson
1	.964[a]	.930	.926	96.696	.888

a．Predictors：(Constant), t

b．Dependent Variable：y

表 14.13　回归系数

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	132.411	41.677	.964	3.177	.005
	t	50.654	3.040		16.665	.000

a．Dependent Variable：y

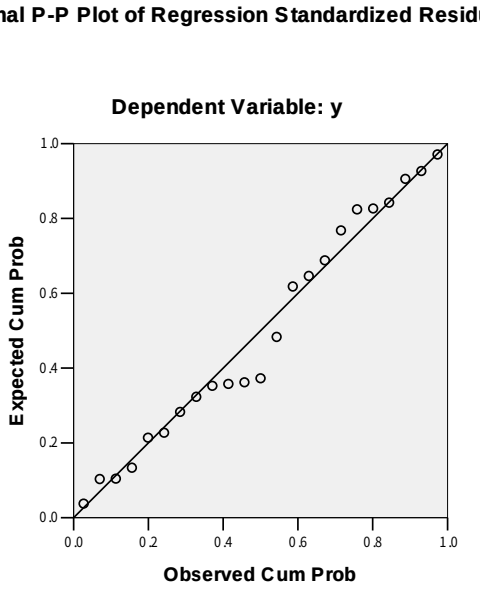


图 14.17　原始序列和预测序列随时间变化的趋势线图

表 14.12、表 14.13 和图 14.17 为获取的部分线性回归结果。表 14.12 列出了模型决定系数和 Durbin-Waterson 统计量等结果；表 14.13 给出模型参数估计结果；图 14.17 为原始序列和预测序列随时间变化的趋势线图。此模型的决定系数为 0.930，相当高，模型拟合得似乎不错，但是 Durbin-Watson 统计量为 0.888，这个数值提示残差有很强的自相关性。进一步看实际值和预测值的线图，发现预测误差在异端时间连续为正，而另一段时间连续为负，特别在序列末端更是如此，这种误差结构是不理想的，所以模型有改进的必要。

下面使用自回归模型进行分析，为了便于和普通回归比较，首先使用 Prais-Winsten 法对数据进行分析。具体操作为，依次单击菜单“ 分析→时间序列→自回归 ”命令，在“ 自回归 ”对话框中选入 *y* 作为因变量，*t* 作为自变量，在“ 方法 ”选项组中选择“ Prais-Winsten ”，如图 14.18 所示。

单击图 14.18 对话框上的“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
* Autoregression.
   TSET PRINT=DEFAULT CNVERGE=.001 CIN=95 NEWVAR=ALL .
```

```
PREDICT THRU END.
AREG y WITH t
/METHOD=PW
/CONSTANT
/RHO=0
/MXITER=10.
```

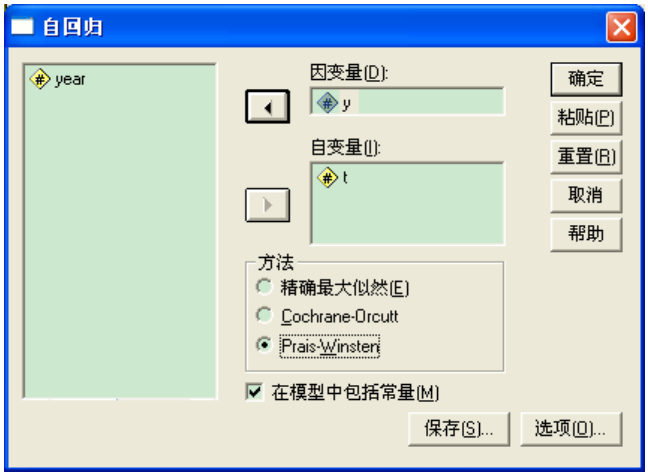


图 14.18 “ 自回归 ” 对话框（选择 Prais-Winsten 方法）

设置完毕后，单击 “ 确定 ” 按钮，执行自回归分析操作。得到其结果分析如表 14.14，表 14.15，表 14.16，表 14.17，表 14.18 和表 14.19 所示。

表 14.14　自相关系数

Autocorrelation Coefficient

Rho (AR1)	Std. Error
.475	.197

The Prais-Winsten estimation method is used.

表 14.14 为自相关系数表，Rho 为一阶自相关系数（实际上就是自回归系数）估计值，为 0.475，自相关系数的标准差（Std.Error）为 0.197。

表 14.15　模型拟合摘要

Model Fit Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin- Watson
.929	.864	.850	85.560	1.426

The Prais-Winsten estimation method is used.

表 14.15 为模型拟合指标摘要。主要的 Durbin-Watson 统计量为 1.426，较线性回归模型的 0.888 更接近 2，这说明残差自相关问题得到结果。

表 14.16　方差分析表

ANOVA

	Sum of Squares	df	Mean Square
Regression	926913.3	1	926913.326
Residual	146410.1	20	7320.504

The Prais-Winsten estimation method is used

表 14.16 给出了方差分析表。

表 14.17　回归系数

Regression Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig
	B	Std. Error	Beta		
T	51.507	4.577	.929	11.253	.000
(Constant)	132.196	63.945		2.067	.052

The Prais-Winsten estimation method is used.

表 14.17 为回归系数参数估计结果。

本例最终得到的模型方程为：

$$\hat{Y}_t = 132.196 + 51.507T_t + 0.475Y_{t-1}$$

本例数据也可以使用“精确最大似然”估计，结果如表 14.18 和表 14.19 所示。

表 14.18 “精确最大似然估计”的残差诊断表

Residual Diagnostics

Number of Residuals	23
Number of Parameters Residual df	1
Adjusted Residual Sum of Squares	20
Residual Sum of Squares	147050.9
Residual Variance	196353.4
Model Std. Error	7237.244
Log-Likelihood	85.072
Akaike's Information	-133.495
Criterion (AIC)	272.991
Schwarz's Bayesian Criterion (BIC)	276.397

表 14.19 “精确最大似然估计”的参数估计

Parameter Estimates

	Estimates	Std Error	t	Approx Sig
Rho (AR1)	.552	.209	2.647	.015
Regression Coefficients t	51.756	5.130	10.089	.000
Constant	132.474	72.151	1.836	.081

Melard's algorithm was used for estimation.

注释：表 14.18 为“精确最大似然估计”的参数估计。其中 Estimates 为估计值，Std Error 为标准误差，t 为 t 检验统计量的观察值，Approx Sig 为自由度的估计熵（该值接近 0，说明其对应的属性可以加入模型）。Rho（AR1）为 AR1 模型中滞后一阶的误差项前面的系数，Regression Coefficients t 为 t 检验分布的衰退系数，Constant 为常数项系数。

14.5 ARIMA 模型

ARIMA 建模实际上包括 3 个步骤，即模型识别阶段、参数估计和检验阶段以及预测应用阶段，其中，前两个阶段可能需要反复进行。

ARIMA 模型的识别就是判断 p、d、q 包括 sp、sd、sq 的阶，主要依靠自相关函数（ACF）和偏自相关函数（PACF）图来初步判断和估计。一个识别良好的模型应该有两个要素：一是模型的残差为白噪声序列，需要通过残差的白噪声检验；二是模型参数的简约性和拟合优度指标的优良性（如对数似然函数值较大，AIC、BIC 较小）方面取得平衡，还有一点需要注意的是模型的形式应该易于理解，比如说，一些长期趋势可以表达为时间的函数，也可以用 1~2 阶差分处理；同样地，有些季节性波动既可以用正弦或余弦函数进行拟合，也可以使用季节性差分处理。如果建立的是回归模型，欲分析各种因素对相应序列的影响，过多使用差分虽然也能很好地拟合模型，但模型参数的实际意义则不明确。所以，如果能通过某种变换或函数拟合使得模型达到平稳，最好不用差分。当然如果仅仅用于预测，使用差分使序列平稳化比使用函数拟合要方便得多。

ARIMA 过程提供建立 Box-Jenkins 的时间序列模型，本过程可以对带 ARIMA 误差的回归方程建模。本过程也可用于建立乘积型季节性模型。

ARIMA 过程的操作非常简单，和 14.4 节的自回归过程类似，对话框行也基本相同，只是多了几个参数设置（如图 14.19 的“ ARIMA 模型 ”对话框，可以通过单击菜单“ 分析→预测→ARIMA 模型 ”命令得到），如果需要（例如，将指数趋势化为线性趋势，处理方差不齐的数据等），可以做自然对数变换。当需要更复杂的数据交换时，可以用“ 转换 ”功能预先处理后再用本过程建模。“ 自变量 ”列表框用于输入解释序列或自变量，可以是一个变量也可以是多个变量，此处如果不选，则拟合 1 个纯粹的 ARIMA 模型；如果选择，则拟合带 ARIMA 误差的回归模型。

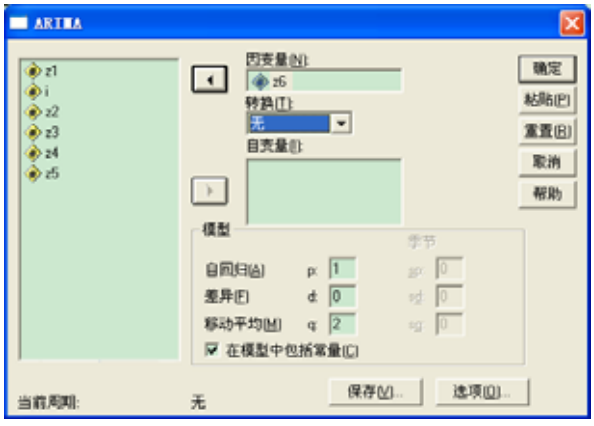


图 14.19 “ ARIMA ”模型对话框

“ 模型 ” 选项组用于输入模型参数细节，*p* 为自回归的阶，*d* 为差分的阶，*q* 为移动平均（也称为 “ 滑动和 ”）的阶。在 “ 季节 ” 选项组下，*sp*、*d*、*sq* 分别填写季节性自回归、差分和移动平均的阶，如果填写此处（填入大于 0 的整数），则得到乘积型季节性 ARIMA 模型，对季节性进行建模的前提是数据中时间周期已经定义。“ 模型 ” 选项组最后的选项询问用户常数项是否需要进入模型，如果有理由认为模型常数项为零，则不选，一般需要选择此项。

在图 14.19 所示对话框中，单击 “ 保存 ” 和 “ 选项 ” 按钮可以得到相应的附加对话框，这里的附加对话框内容和图 14.12、图 14.13 内容基本相同，此处不再赘述，读者可以参考相关说明。

虽然 ARIMA 模块并不复杂，但是建立一个好的或较优的 ARIMA 模型却非易事。一方面，ARIMA 模型的形式不是唯一的（数学上可以证明），例如，一个 *p* 阶自回归模型理论上可以用无穷阶的移动平均模型来精确刻画。另一方面，目前尚无任何程序帮助用于自动选择 *p*、*d*、*q* 的阶，这些需要研究者去自行判断。

【示例 14.5】ARIMA 模型示例。

试对模拟数据（如图 14.19 所示）的 6 个序列分别做 ARIMA 模型的初步识别。从图 14.19 可以看出，数据文件包括 $z_1 \sim z_6$ 共 6 个模拟的随机序列，*i* 为时间顺序变量。首先进行模型识别，对此 6 个序列分别做 ACF 和 PACF 图。

01 打开时间序列图形对话框，将待做图的变量选入（如图 14.20 所示的 “ 自相关 ” 对话框，可以通过依次单击菜单 “ 分析→预测→自相关 ” 命令得到），在 “ 显示 ” 选项组中选择 “ 自相关 ” 和 “ 偏自相关 ”。



图 14.20 “ 自相关 ” 对话框

02 单击图 14.20 对话框上 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

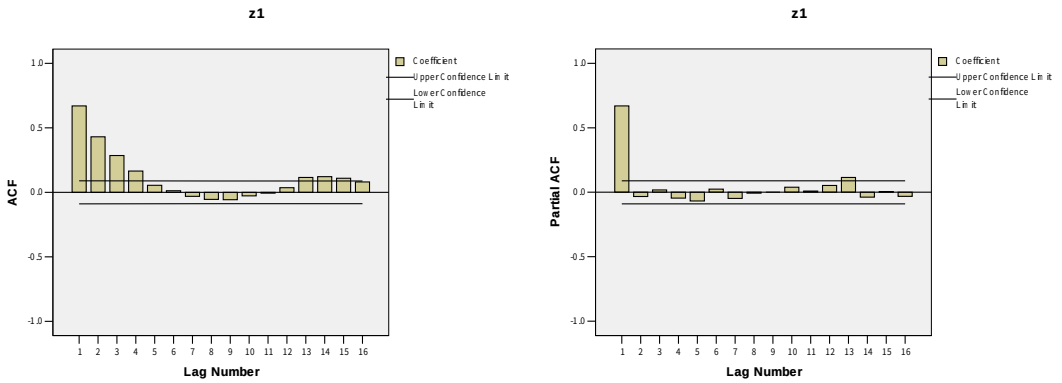
```
ACF
    VARIABLES= z2 z3 z4 z5 z6 z1
    /NOLOG
    /MXAUTO 16
    /SERROR=IND
```

```
/PACF.
```

03 设置完毕后，单击“确定”按钮，执行相关操作。

ACF 和 PACF 图为相关系数函数图，横坐标为时间间隔或称时滞，纵坐标为相关系数，取值为-1~1 之间，图中的柱子标示在一定时滞下自相关系数的值，图中两条横线为相关系数假设检验参考标准线，在两线之间的相关系数无统计学意义，超出两线间的柱子所代表的相关系数有统计学意义。

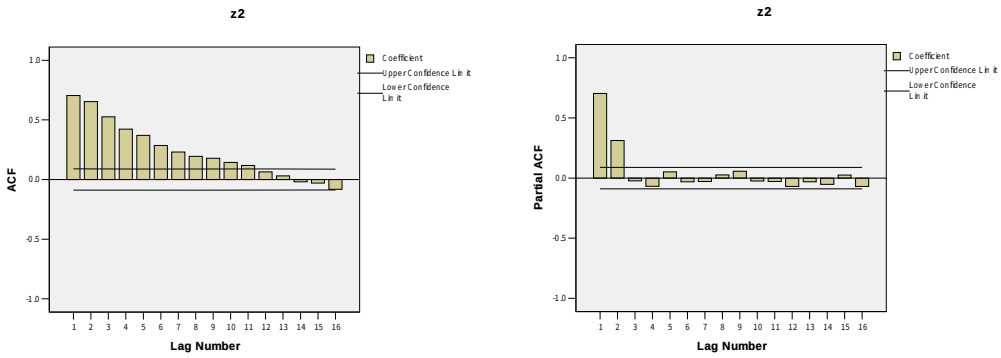
图 14.21 为序列 z_1 的 ACF 图和 PACF 图。从图中可以看出，ACF 图呈拖尾衰减，PACF 一步截尾，可判断为平稳序列，识别为 AR1 模型，即 $p=1$、$d=0$、$q=0$ 的 ARIMA（1，0，0）模型。



（a）序列 z_1 的 ACF 图　　（b）序列 z_1 的 PACF 图

图 14.21　序列 z_1 的 ACF 图和 PACF 图

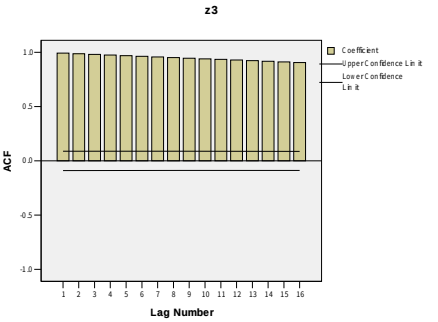
图 14.22 为序列 z_2 的 ACF 图和 PACF 图。从图中可以看出，ACF 图呈拖尾衰减，PACF 两步截尾，可判断为平稳序列，识别为 AR2 模型，即 $p=2$、$d=0$、$q=0$ 的 ARIMA（2，0，0）模型。



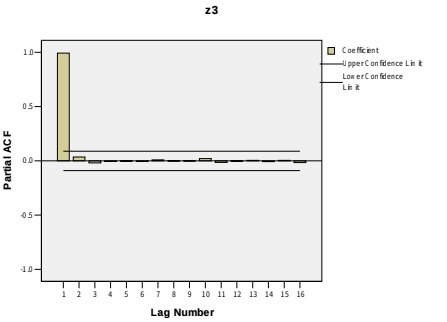
（a）序列 z_2 的 ACF 图　　（b）序列 z_2 的 PACF 图

图 14.22　序列 z_2 的 ACF 图和 PACF 图

图 14.23 为序列 z_3 的 ACF 图和 PACF 图。从图中可以看出，ACF 图呈拖尾衰减缓慢，为非平稳序列特点，PACF 一步截尾，尚无法识别，需要将序列平稳化处理后再进行判断。



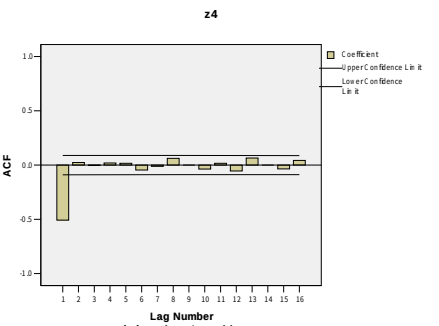
(a) 序列 z_3 的 ACF 图



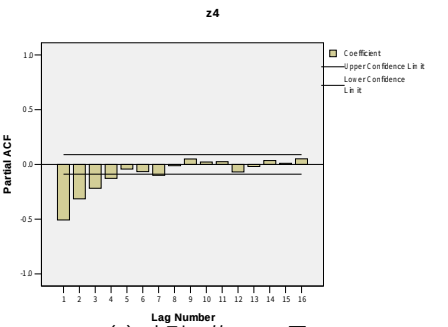
(b) 序列 z_3 的 PACF 图

图 14.23　序列 z_3 的 ACF 图和 PACF 图

图 14.24 为序列 z_4 的 ACF 图和 PACF 图。从图中可以看出，ACF 图一步截尾，PACF 呈拖尾衰减，可判断为平稳序列，识别为 MA1 模型，即 $p=0$、$d=0$、$q=1$ 的 ARIMA(0,0,1) 模型。

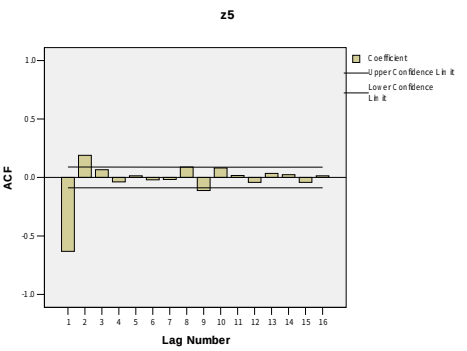


(a) 序列 z_4 的 ACF 图

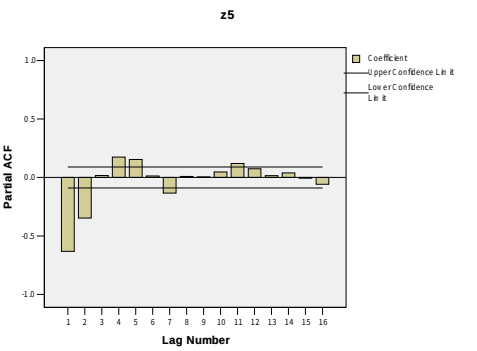


(b) 序列 z_4 的 PACF 图

图 14.24　序列 z_4 的 ACF 图和 PACF 图



(a) 序列 z_5 的 ACF 图

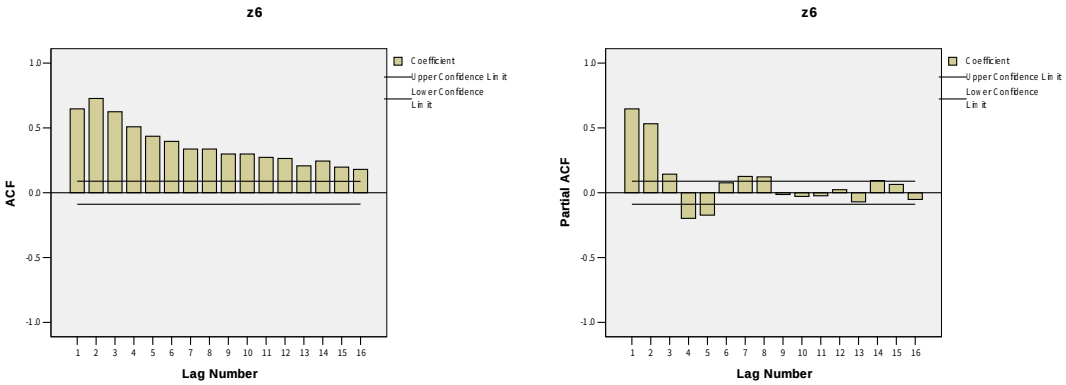


(b) 序列 z_5 的 PACF 图

图 14.25　序列 z_5 的 ACF 图和 PACF 图

图 14.25 为序列 z_5 的 ACF 图和 PACF 图。从图中可以看出，ACF 图两步截尾，PACF 呈拖尾衰减，可判断为平稳序列，识别为 MA2 模型，即 $p=0$、$d=0$、$q=2$ 的 ARIMA(0,0,2) 模型。

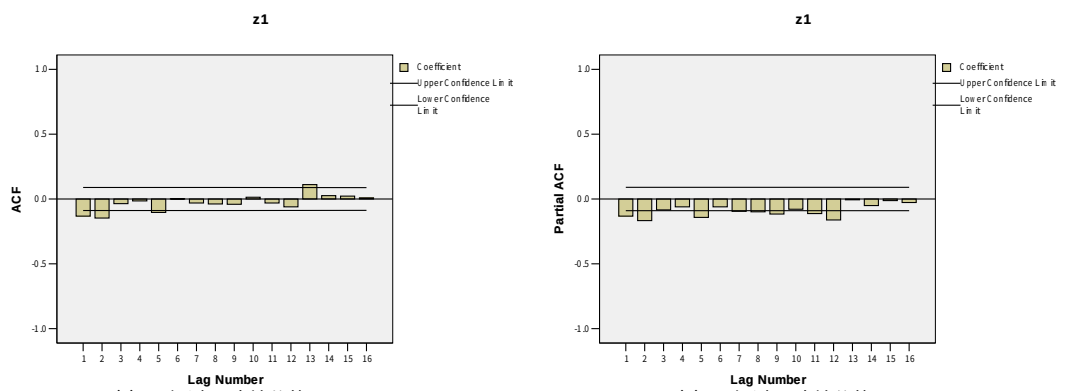
图 14.26 为序列 z_6 的 ACF 图和 PACF 图。从图中可以看出，ACF 图和 PACF 图均呈拖尾衰减，可判断为平稳序列，识别为混合模型，即 $p=$？、$d=0$、$q=$？的 ARIMA（？，0，？）模型。



(a) 序列 z_6 的 ACF 图　　(b) 序列 z_6 的 PACF 图

图 14.26　序列 z_6 的 ACF 图和 PACF 图

因为序列 z_3 为非平稳序列，需要对其进行平稳化处理后再进行识别，最常用的方法是做差分处理，即在选中图 14.20“转换”选项组的“差分”复选框，并在其后的文本框中输入 1，做一阶差分后再进行识别。由结果（如图 14.27 所示）可见，差分序列的 ACF 图呈两步截尾，而 PACF 图呈一步截尾，初步识别为 AR1 模型（ACF 当成拖尾处理，也就是当 ACF 和 PACF 都貌似截尾时，把尾巴长的当拖尾处理）。由于原始序列已经做了一阶差分处理，所以，原始序列 z_3 识别为 ARIMA（1，1，0）。



(a) z_3 经过一阶差分的 ACF 图　　(b) z_3 经过一阶差分的 PACF 图

图 14.27　z_3 经过一阶差分后的 ACF 图和 PACF 图

模型初步识别结束后，下一步进行参数估计和模型诊断，对于特征明显的序列，一次识别就可确定模型的阶数，并通过模型诊断，而混合模型则需要反复尝试。

利用上面的识别结果，对序列 z_3 进行参数估计和模型诊断。

序列 z_3 已经明确识别为 ARIMA（1，1，0），参数估计只需要调用 ARIMA 过程，在 ARIMA 过程的对话框中（如图 14.20 所示）将 z_3 选入“因变量”列表框，在参数 p、d、q 栏中依次填入 1、1、0。

04 单击图 14.20 对话框上“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
* ARIMA.
TSET PRINT=DEFAULT CIN=95 NEWVAR=ALL .
PREDICT THRU END.
ARIMA z3
    /MODEL=( 1 1 0 )CONSTANT
    /MXITER= 10
    /PAREPS= .001
    /SSQPCT= .001
    /FORECAST= EXACT .
```

05 设置完毕后，单击“确定”按钮，执行相关操作。主要输出结果如下。

- 模型的描述信息统计表

表 14.20　模型描述信息统计表

Model Description[a]

Model Name	MOD_11
Dependent Series	z3
Transformation	None
Constant	Included
AR	1
Non-Seasonal	1
Differencing	
MA	None

Applying the model specifications from MOD_11

a．Since there is no seasonal component in the model, the seasonality of the data will be ignored.

表 14.20 为 ARIMA 模型的描述信息统计表，包括：系统按顺序自动生产的模型的名字，这里是 MOD_11，以下指出响应序列为 z_3，未做数据转换，模型包含常数项，自回归阶数为 1，非季节性差分阶数为 1，无移动平均算子（相当于移动平均阶数为 0）。

- 模型残差统计量和拟合优度统计量列表

表 14.21　模型残差统计量和拟合优度统计量列表

Residual Diagnostics

Number of Residuals	499
Number of Parameters	1

续表

Residual df	497
Adjusted Residual Sum of Squares	501.634
Residual Sum of Squares	501.640
Residual Variance	1.009
Model Std. Error	1.005
Log-Likelihood	-709.366
Akaike's Information Criterion (AIC)	1422.731
Schwarz's Bayesian Criterion (BIC)	1431.157

表 14.21 为模型残差统计量和拟合优度统计量列表。

- 参数估计结果

表 14.22　参数估计结果

Parameter Estimates

	Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags AR1	-.343	.042	-8.09	.000
Constant	.242	.034	7.223	.000

Melard's algorithm was used for estimation.

注释：表 14.22 为参数估计结果。其中 Estimates 为估计值，Std Error 为标准误差，t 为 t 检验统计量的观察值，Approx Sig 为自由度的估计熵（该值接近 0，说明其对应的属性可以加入模型）。Non-Seasonal Lags AR1 为 AR1 模型中滞后的非季节性因子项前面的系数，Constant 为常数项系数。

- 残差自相关检验结果

表 14.23 为残差自相关检验的结果，第一列是时滞，第二列是自相关系数，第三列是自相关系数的标准误差，后 3 列分别为检验统计量、自由度和 ρ 值。如果在任何时滞上出现小的 ρ 值（比如说小于 0.05），则认为残差存在自相关性，模型拟合不足，需要改进。如本例，ρ 值都远大于 0.05，可以认为残差为白噪声序列，模型诊断得以通过。

表 14.23　残差自相关检验结果

Autocorrelations

Series：z3					
Lag	Autocorre lation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig[b]
1	-.340	.045	57.885	1	.000
2	.135	.045	67.037	2	.000
3	-.016	.045	67.170	3	.000
4	-.035	.044	67.800	4	.000
5	.018	.044	67.95968.060	5	.000
6	-.014	.044	68.301	6	.000

续表

Series：z3					
Lag	Autocorre lation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig[b]
7	.022	.044	68.525	7	.000
8	-.021	.044	71.624	8	.000
9	.078	.044	72.619	9	.000
10	-.044	.044	73.165	10	.000
11	.033	.044	73.371	11	.000
12	-.020	.044	73.480	12	.000
13	.015	.044	73.483	13	.000
14	.002	.044	73.484	14	.000
15	-.001	.044	73.730	15	.000
16	.022	.044		16	.000

a．The underlying process assumed is independence (white noise)

b．Based on the asymptotic chi-square approximation.

- ACF 图

图 14.28 为 ACF 图，观察 ACF 图，可以很直观地看到各时滞上的自相关系数均无统计学意义，PACF 图没有给出，读者可自行操作本例数据，观察有关结果。

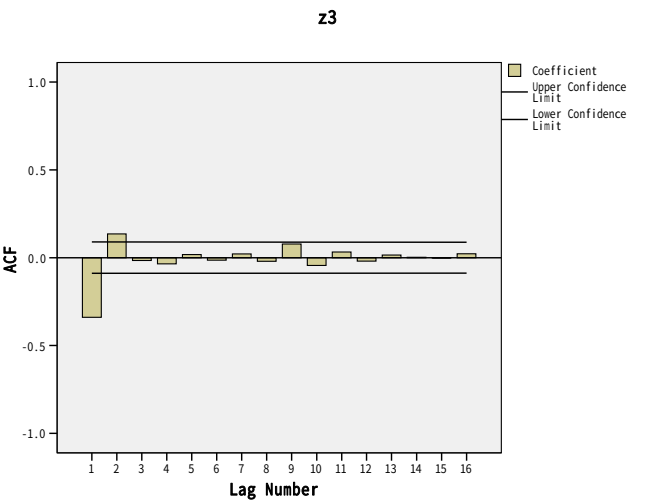


图 14.28　ACF 图

本例得到的最终模型（为方便叙述，令 $Y = z_3$）的数学表达式为

$$(1-\phi_1 B)(1-B)Y_t = u + a_t$$

其中，ϕ_1 自回归系数，u 为常数，a_t 为白噪声。将参数代入，得：

$$(1+0.343_1 B)(1-B)Y_t = 0.242 + a_t$$

进一步简化，得：

$$Y_t + 0.343_1 BY_t - BY_t - 0.343B^2Y_t = 0.242 + a_t$$
$$Y_t = 0.242 + 0.657Y_{t-1} + 0.343Y_{t-2} + a_t$$

注意：为了能看懂参数估计和正确写出 ARIMA 模型的方程式，读者必须熟悉 ARIMA 模型的因式表达，对此不熟悉的读者应仔细阅读本章前面内容。这里还可以发现，ARIMA（1，1，0）通过恒等变换，居然变成了一个 AR2 模型。由两参数模型（1 个自回归系数和 1 个常数）变成了三参数模型（2 个自回归系数和 1 个常数）。从这里也可以发现，ARIMA 模型的形式不是唯一的。细心的读者会发现，AR2 形式模型的两个自回归系数并不能自由取值，存在所谓参数冗余的现象。

对示例 14.5 建模再讨论：比如，有理由认为非平稳随机序列汇总有随时间变化的长期趋势，z_3 是模拟数据，我们知道 z_3 中有时间变量 i 的线性函数成分，所以可以先以 z_3 为因变量，以 i 为自变量建立线性回归方程，然后再去对此回归方程的残差拟合 ARIMA 模型。当知道残差模型机构后，直接拟合 ARIMAX 模型，最终结果为：

$$Y_t = -9.691 + 0.655Y_{t-1} + 0.341Y_{t-1} + 0.243i_t + a_t$$

其中，$Y = z_3$，a_t 为白噪声。此结果请读者自行证明。

再次利用上面的识别结果，对序列 z_6 建立 ARIMA 模型。

在上面的分析中，已经得出结论：z_6 为拟合混合模型，ARIMA（？，0，？）模型，所未知的是 p、q 的阶。对此类问题的思路是，通过尝试法从简单到复杂建立模型，直到模型残差通过白噪声检验；在残差检验通过的模型中，选择参数简单的模型，同时兼顾拟合优度统计量（常用 AIC 和 BIC，两者越小，模型拟合越好）。

经验发现，对于多数数据，p、d、q 取 2 或以下都能满足拟合混合模型的需要。对于这一步的结果，在此就不给出，具体操作与前面的示例类似，读者可以自行分析。

14.6 周期性分解

周期性分解模型具有等价的 *ARIMA（p，d，q）×（P，D，Q）*s 表示形式，并有相应的参数估计和假设检验方法。SPSS 的周期性分解过程比较简单，只能处理确定性的季节成分和只有基本的参数估计结果，没有提供有关的假设检验内容。

传统的季节分量模型将随机序列主观地分解成 3 个组成部分，或称为 3 个分量，即“趋势分量”、“季节分量”和“随机波动”，趋势分量使用多项式拟合，季节分量使用傅里叶变换来估计。其数学表达式为：

$$Y_t = f(T_t, S_t, I_t) \qquad (14.25)$$

其中，T_t 代表长期趋势（可以是线性趋势，也可以是周期性波动或长周波动），S_t 为季节因子（幅度和周期固定的波动，日历效应为常见的季节因子），I_t 为随机波动（可视为误差）。函数 f 有加法和乘法两种，常用乘法模型。

季节分量模型在应用过程中有两大缺点：其一是人为地将随机序列分解成 3 个固定的

成分不一定科学；其二是当有新的数据加入序列后，所有的分量需要重新估计。

季节分量模型要求无缺失数据，在处理前数据已经由 SPSS 系统定义好时间变量并指定周期。

打开“ 周期性分解 ”对话框（如图 14.29 所示的“ 周期性分解 ”对话框，可以通过依次单击菜单“ 分析→预测→季节性分解 ”命令得到），将待分析的序列变量名选入“ 变量 ”列表框。在“ 模型类型 ”选项组中选择模型，该对话框提供两种，即“ 乘法 ”模型（系统默认选项）和“ 加法 ”模型；“ 移动平均权重 ”有两个单选项：“ 所有点相等 ”和“ 结束点按 0.5 加权 ”，如果周期为奇数时间单位，则不必选择此项；如果需要将所有结果在输出窗口显示，则需要选中“ 显示对象删除列表 ”选项。

“ 周期性分解：保存 ”对话框选择是否将结果以新变量的形式添加到原始数据集中，如图 14.30 所示。本例采用默认选项。



图 14.29 “ 周期性分解 ”对话框

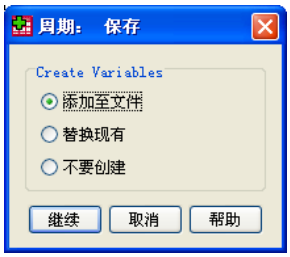


图 14.30 “ 周期：保存 ”对话框

【示例 14.6】周期性分解示例。

表14.24为某公司连续144个月的月度销售量记录，变量为sales。试用周期性分解模型分析此数据。

表 14.24　某公司连续 144 个月月度销售记录数据

日　　期	sales	日　　期	sales
1978-9-1	112	1979-8-1	118
1978-10-1	118	1979-9-1	115
1978-11-1	132	1979-10-1	126
1978-12-1	129	1979-11-1	141
1979-1-1	121	1979-12-1	135
1979-2-1	135	1980-1-1	125
1979-3-1	148	1980-2-1	149
1979-4-1	148	1980-3-1	170
1979-5-1	136	1980-4-1	170
1979-6-1	119	…	…
1979-7-1	104		

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 打开周期性分解对话框和选入变量。

依次单击菜单“ 分析→预测→季节性分解 ”命令，打开“ 周期性分解 ”对话框。在“ 变量 ” 列表框中选择变量“ sales ”；在 “ 模型类型 ” 选项组中选择 “ 乘法 ” 选项；在 “ 移动平均权重 ” 选项组中选择 “ 终点按 0.5 加权 ”。

03 设置。

单击 “ 保存 ” 按钮，打开 “ 周期：保存 ” 对话框，选择 “ 添加至文件 ” 选项。

04 单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
* Seasonal Decomposition.
TSET PRINT=DEFAULT NEWVAR=ALL .
SEASON
    /VARIABLES=sales
    /MODEL=MULTIPLICATIVE
    /MA=CENTERED.
```

05 设置完毕后，单击 “ 确定 ” 按钮，执行周期性分解操作。

执行完毕后，得到分析结果变量，这样我们就可以通过序列图对结果进行分析比较了。关于 “ 序列图 ” 方面的操作，不是本章的重点，在此不做介绍，对此不熟悉的读者可以参阅本书其他章节知识。得到其结果分析如下所示。

- 原始序列和校正了季节因子作用的序列图

图 14.31 为原始序列和校正了季节因子作用的序列图。蓝线为原始序列，体现了销售量呈年度周期震荡增长的特征。绿线为校正了的月度效应序列，在 12 年里呈稳步增长的态势。

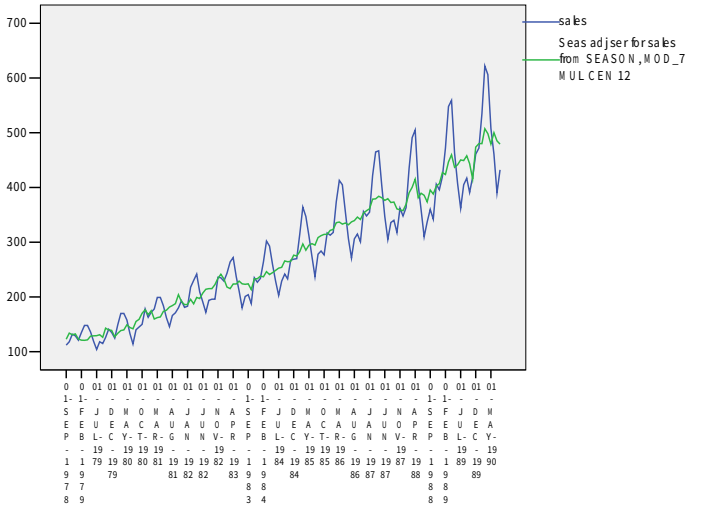


图 14.31　原始序列和校正了季节因子作用的序列图

- 季节因子图

图 14.32 为季节因子图，呈 12 个月周期的规则波动；可发现一年中，6~9 月公司的销售量较大，其他时间相对较少，1~2 月为销售淡季。

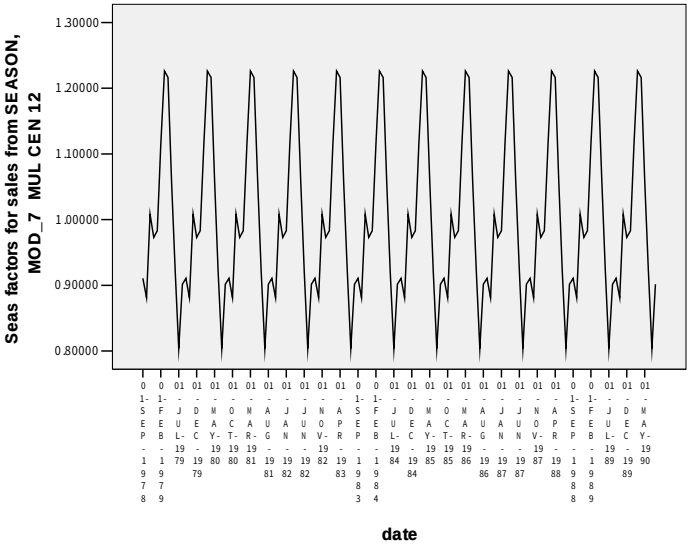


图 14.32　季节因子图

- 趋势成分图

图14.33为趋势成分图。趋势成分图反映公司销售量在12年里呈增长的态势，前8年基本平稳增长，后4年虽然在总体上维持了前8年增长的态势，但增长过程波动较大。

- 随机波动成分图

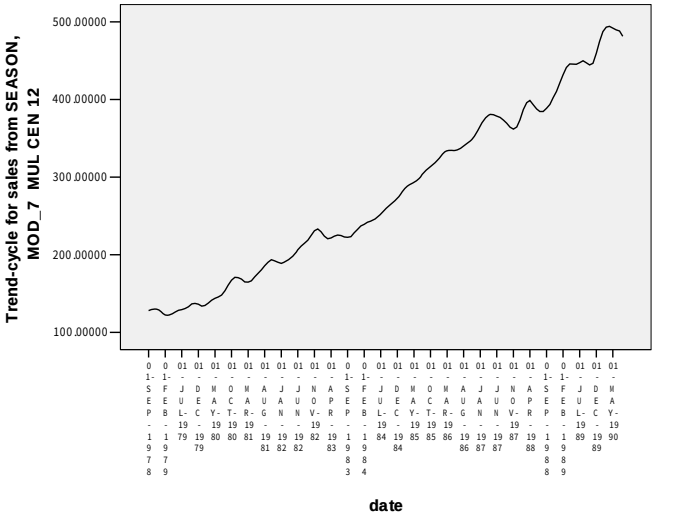


图 14.33　趋势成分图

图 14.34 为随机波动成分图，可能含有模型未能解释的因素。

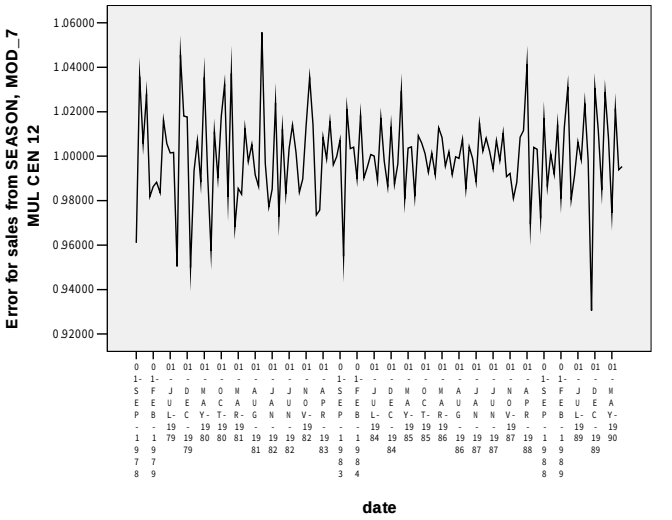


图 14.34　随机波动成分图

14.7　本章小结

本章给出了时间序列的基本原理和几种常用的时间序列分析方法。时间序列分析在高级统计分析中已经得到了极为广泛的应用，尤其是在经济学、医学和社会学研究领域已经成为了一种必需的分析方法。

SPSS 的时间序列分析模块给出了 4 种时间序列分析方法。读者学习时，要重点掌握和理解时间序列分析的基本思想和 SPSS 操作中输出结果所包含的内容。

第15章　生存分析

在科学研究中，可通过随访（Follow Up）来研究事物的变化发展规律，以获取有关生存的信息。随访资料的特点是：

- 需随访一段时间，由于研究时间较长，难免出现研究个体失访等原因而退出研究的现象，形成截尾数据；
- 数据分布类型复杂，一般生存时间数据的分布呈正偏态分布。

正是由于这些缺点，研究分析时不仅需要考虑某种结局（如有效、治愈、死亡等），还需要考虑出现这些结局所经历的时间长短。生存分析通常采用寿命表分析、Kaplan-Meier 分析等非参数方法计算与比较单因素生存率；采用 Cox 比例风险回归模型等半参数方法考虑多个因素对生存情况的影响。

在 SPSS 中进行聚类分析，可以利用“分析→生存函数”命令实现，菜单内容如图 15.1 所示。

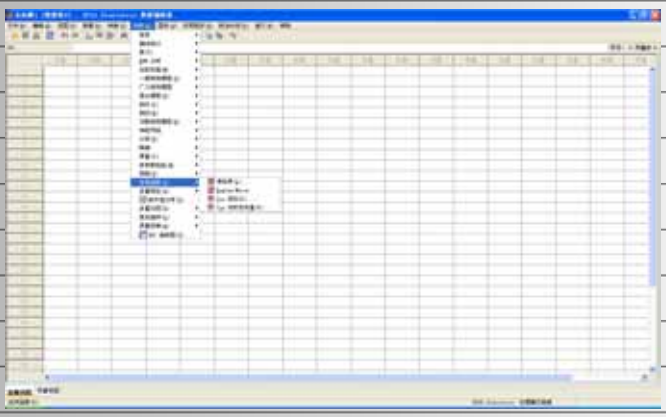


图 15.1　SPSS 生存分析模块

本章学习要点：

- 生存分析概述；
- 寿命表分析；
- Kaplan-Meier 分析；
- Cox 回归分析。

15.1 生存分析概述

本节主要向读者介绍生存分析的常用术语。

（1）生存时间（Survival Time）

生存时间可定义为从某起始事件到达某终点事件所经历的时间跨度。起始时间和终点事件根据研究目的和专业知识在设计阶段确定。起始事件如疾病的确诊、某种处理（治疗）的实施等，终点事件可以是某种疾病的发生、某种处理（治疗）的反应、病情的复发或死亡等，又称失效事件（Failure Event）。生存时间常用符号 t 表示。

（2）完全数据（Complete Data）

在随访期间，随访对象发生了失效事件，即观察到随访对象出现了我们所规定的结局，该观察对象所提供的关于生存时间的信息是完整的，这种生存时间数据称为完全数据。例如，某研究观察了 10 名进行输卵管结扎术后的妇女经峡部—峡部输卵管吻合术后的受孕时间（月）分别为 2、3、3、4、4、7、8、10、13 和 15，这就是一组按由小到大的顺序整理过的完全数据。

（3）截尾数据（Censored Data）

在实际追踪观察中，由于某种原因无法知道观察对象的确切生存时间，这种生存时间数据称为截尾数据。例如，有 10 名行输卵管结扎手术的妇女经壶腹—壶腹部吻合术后的受孕时间（月）分别为：4、5、5、6、7、10^+、14^+、20^+、31^+和 44。这就是一组按由小到大的顺序整理过的数据，其中带有“+”号的数字为截尾数据。产生截尾数据的原因大致有以下两个：

- 观察对象失访

例如，因搬迁而失去联系或中途退出试验，或因其他的与本研究无关的原因死亡（或失败）而未能观察到规定的终点。终止随访时间为失访时间（或死亡时间）。

- 观察对象的生存期超过了研究的终止期

例如，研究计划规定只对病人随访 4 年，但有的病人的生存期超过了 4 年，或者由于病人进入研究的时间较晚，虽然对他的随访期未满 4 年，但已到了研究的截止时间。

不论截尾数据的产生原因为何，截尾生存时间的计算均为起始事件至截尾点所经历的时间。常见的右截尾（Right Censoring）表示准确的生存时间长于截尾时间。截尾数据常在其右上角标记“+”号。

（4）生存率或生存函数（Survival Function）

令 T 表示生存时间，生存率或生存函数表示观察对象活过时间 t 的概率，又称累积生存函数（Cumulative Survival Function），符号为 $S(t)$。

$$S(t)=P(T>t),0\leqslant S(t)\leqslant 1$$
$$\hat{S}(t)=\frac{\text{活过时间}t\text{的观察例数}}{\text{观察总例数}} \tag{15.1}$$

以生存时间为横轴，生存率为纵轴，将各个时间点所对应的生存率连接在一起的曲线图称为生存曲线（Survival Curve）。

（5）风险函数（Hazard Function）

风险函数又称危险函数，表示一个生存到时间 t 的观察对象，从 t 到 $t+\Delta t$ 这一区间内死亡的概率极限，常用 $h(t)$ 表示。其计算公式为：

$$h(t)=\lim(\Delta t\to\infty)\frac{t\text{时间生存者死亡区间}(t,t+\Delta t)\text{的概率}}{\Delta t} \tag{15.2}$$

式（15.2）是风险函数 $h(t)$ 的定义式。在实际工作中，风险函数可用式（15.3）来估计：

$$h(t)=\frac{\text{死亡区间}(t,t+\Delta t)\text{人数}}{t\text{时尚存人数}\times\text{该区间所含单位时间数}}=f(t)/S(t) \tag{15.3}$$

（6）中位生存时间和平均生存时间

中位生存时间（Median Survival Time）又称半数生存期，表示恰有 50%的个体尚存活的时间，即生存曲线上纵轴 50%所对应横轴的生存时间。

平均生存时间（Mean Survival Time）则表示生存曲线下的面积。

15.2　寿命表法

对于生存资料，首先需要给出各时间点上生存函数的估计值，方法之一是采用寿命表法（Lift-Table Method，简称 LT 法）。寿命表法适用于区间数据，通过计数落入时间区间 $[t,t+\Delta t]$ 内的失效和截尾的观察例数来估计该区间上的死亡概率，然后用该区间及其之前的各区间上的生存概率之积来估计 $S(t)$。

寿命表法适用于样本含量大的资料，在 SPSS 中，可由以下步骤实现。

- 选择寿命表分析对话框

依次单击菜单“分析→生存函数→寿命表”命令，打开“寿命表”对话框，如图 15.2 所示。

- 选择变量

在左侧变量列表框中选择变量后，单击 按钮，可以将其移动到“时间”、“状态”、“因子”或“按因子”列表框中，如图 15.2 所示。

在图 15.2 所示对话框“显示时间间隔”选项组中可以输入欲输出的生存时间范围及组距。在“0 到”后面的文本框中输入生存时间上限，在“步长”后面的文本框中输入生存时间的组距。

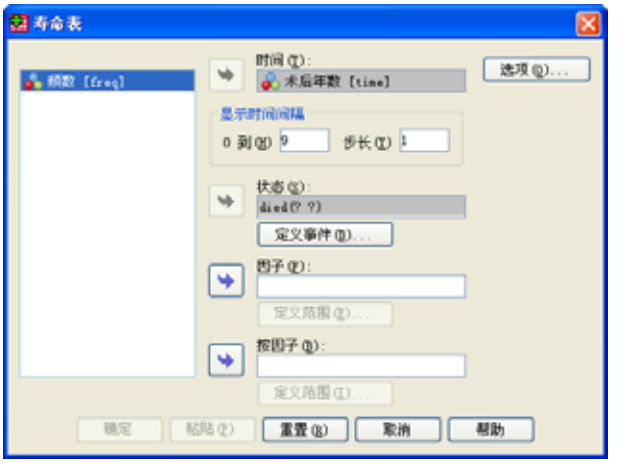


图 15.2 “ 寿命表 ” 对话框

- “ 选项 ” 选项

在图 15.2 所示对话框中，单击 “ 选项 ” 按钮，打开 “ 寿命表：选项 ” 对话框，选择分层聚类中距离的类别。如图 15.3 所示。

在图 15.3 所示的 “ 寿命表：选项 ” 对话框中，有 “ 寿命表 ”、“ 图 ” 和 “ 比较第一个因子的水平 ” 三个部分。

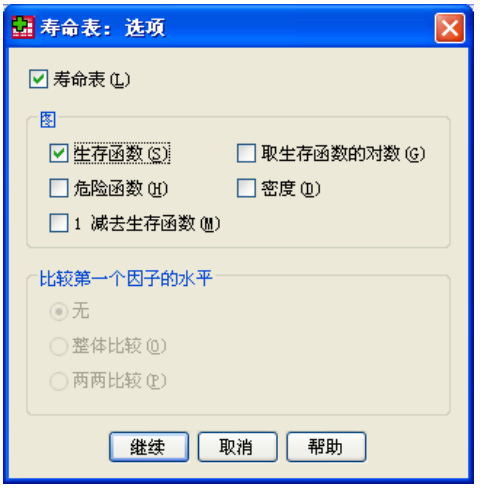


图 15.3 “ 寿命表：选项 ” 对话框

“ 寿命表 ” 选项表明可以输出寿命表，为系统默认选项。

“ 图 ”选项组总共可以输出 5 种（可复选）绘图，包括：“ 生存函数 ”、“ 危险函数 ”、“ 1 减去生存函数 ”、“ 取生存函数的对数 ” 和 “ 密度 ”。

“ 比较第一个因子的水平 ” 选项组给出第一个因子不同水平的比较。包括 “ 无 ”、“ 整

体比较 ” 和 “ 两两比较 ” 三个单选按钮。

【示例 15.1】寿命表分析。

现有 346 例大肠癌患者的随访资料，如表 15.1 所示。试描述其生存情况。

表 15.1　346 例大肠癌患者术后生存情况

术后年数	0~	1~	2~	3~	4~	5~	6~	7~	8~	9~
期间死亡人数	88	80	59	36	12	8	4	7	5	0
期间删失人数	2	1	3	15	8	9	3	3	1	2

本资料是以频数表的方式整理的，因此在分析前需指定频数变量 freq ；分组方式为 0~1 年、1~2 年等，为了便于录入，用组段的起始年数表示该组段 ；结局 died=1 表示死亡，died=0 表示删失。重新整理后的数据如表 15.2 所示。

表 15.2　整理后的 346 例大肠癌患者术后的生存状况

术后年数（time）	频数（frreq）	结局（died）	术后年数（time）	频数（freq）	结局（died）
0	88	1	5	8	1
0	2	0	5	9	0
1	80	1	6	4	1
1	1	0	6	3	0
2	59	1	7	7	1
2	3	0	7	3	0
3	36	1	8	5	1
3	15	0	8	1	0
4	12	1	9	0	1
4	8	0	9	2	0

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 变量加权。

本资料是以频数表的方式整理的，在估计生存函数时，需要先进行变量加权。

依次单击菜单 “ 数据→加权个案 ” 命令，打开 “ 加权个案 ” 对话框，如图 15.4 所示。

选择 “ 加权个案 ”，将 “ 频数（freq）” 选入 “ 频率变量 ” 列表框。单击 “ 粘贴 ” 按钮得到的 Syntax 命令语句程序如下：

```
WEIGHT
   BY freq .
```

设置完毕后，单击 “ 确定 ” 按钮，执行加权个案操作。

03 寿命表分析。

依次单击菜单 “ 分析→生存函数→寿命表 ” 命令，打开 “ 寿命表 ” 对话框。将 “ 术后

年数（time）”选入“时间”列表框；将“结局（died）”选入“状态”列表框。在“显示时间间隔”选项组左右两个文本框中分别输入 9 和 1。

04 设置。

单击“选项”按钮，打开“寿命表：选项”对话框，依次选择“寿命表”和“生存函数”选项。

单击“定义事件”按钮，打开“寿命表：为状态变量定义事件”对话框，选择“单值”选项，在文本框中输入 1，如图 15.5 所示。



图 15.4 “加权个案”对话框

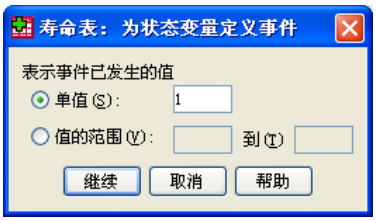


图 15.5 “寿命表：为状态变量定义事件”对话框

05 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
SURVIVAL
TABLE=time
    /INTERVAL=THRU 9 BY 1
    /STATUS=died(1)
    /PRINT=TABLE
    /PLOTS ( SURVIVAL )=time  .
```

设置完毕后，单击“确定”按钮，执行寿命表分析操作。

得到其结果分析如表 15.3 所示。

- 寿命表

表 15.3 寿命表

Life Table[a]

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal gvents	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std.Error of Cumulative Proportion Surviving at End of Interval	Probability Denslty	Std.Error of Probability Density	Hazard Rate	Std.Error of Hazard Rate
.000	346	2	345.000	88	.26	.74	.74	.02	.255	.023	.29	.03
1.000	356	1	255.500	80	.31	.69	.51	.03	.233	.023	.37	.04
2.000	175	3	173.500	59	.34	.66	.34	.03	.174	.021	.41	.05

续表

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal gvents	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std.Error of Cumulative Proportion Surviving at End of Interval	Probability Denslty	Std.Error of Probability Density	Hazard Rate	Std.Error of Hazard Rate
3.000	113	15	105.500	36	.34	.66	.22	.02	.115	.018	.41	.07
4.000	62	8	58.000	12	.21	.79	.18	.02	.046	.013	.23	.07
5.000	42	9	37.500	8	.21	.79	.14	.02	.038	.013	.24	.08
6.000	25	3	23.500	4	.17	.83	.12	.02	.024	.011	.19	.09
7.000	18	3	16.500	7	.42	.58	.07	.02	.049	.016	.54	.20
8.000	8	1	7.500	5	.67	.33	.02	.01	.044	0.17	1.00	.39

表 15.3 是大肠癌病人的寿命表，其中“ The median survival time is 2.07 ”表示中位生存时间为 2.07 年，即术后大肠癌患者死亡人数达到一半的时间为 2.07 年。寿命表中各指标含义的说明如表 15.4 所示。

表 15.4　各指标含义

指 标 名	含　义
Interval Start time	生存时间的组段下限
Number Entering Interval	进入该组段的观察例数
Number Vithdrawing during Interval	进入该组段的删失例数
Number Exposed to Risk	暴露于危险因素的例数，即有效观察例数
Number of Terminal Events	出现失效事件的例数，即死亡（复发、恶化）例数
Proportion Terminating	失效事件比例，即死亡概率
Proportion Surviving	生存概率，等于（1-死亡概率）
Cumulative Proportion Surviving at End of Interval	至本组段上限的累积生存率，由各组的生存概率累积相乘得到
Std.Error of Cumulative Proportion Surviving at End of Interval	累积生存率的标准误差
Probability Density	概率密度，所有个体在时点 t 后单位时间内死亡概率的估计值
Std.Error of Probability Density	概率密度的标准误差
Hazard Rate	风险率，表示活过时点 t 后单位时间内死亡的概率估计值
Std.Error of Hazard Rate	风险率标准误差

- 累积生存率曲线

图 15.6 为累积生存率曲线图，从图中可以看出，随着时间的推移，死亡人数越来越多。其中在开始的 4 年里，死亡人数下降速度较快，随后逐渐趋于平稳。

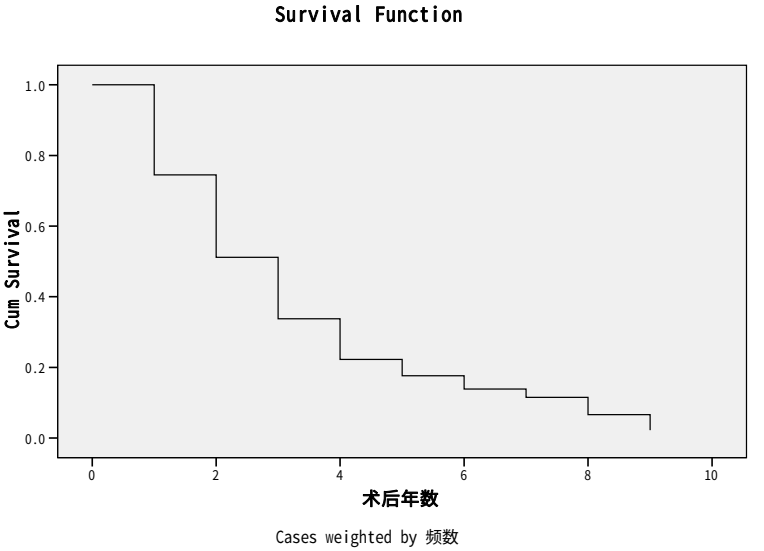


图 15.6　累积生存率曲线

15.3　Kaplan-Meier 分析

由英国科学家 Kaplan 和 Meier 于 1958 年提出的 Kaplan-Meier 分析法（K-M 法），是利用概率乘法定理计算生存率，故又称乘积限法（Product-Limit Method，P-L 法）。Kaplan-Meier 过程适用于小样本或大样本未分组资料生存率的 Kaplan-Meier 分析，生存率估计和组间生存率比较。

（1）生存率的点估计

设 n_{i-1}、n_i、d_i 和 c_i 分别表示活过时间 t_{i-1} 且未在 t_{i-1} 截尾的对象数、期初例数、死亡数和截尾数，则时间 t_i 处的生存率估计为：

$$\hat{S}(t)=(1-\frac{d_1}{n_0})(1-\frac{d_2}{n_1})...(1-\frac{d_i}{n_{i-1}}),i=1,2,...,k \tag{15.4}$$

K-M 估计的各性质如下：

- 要求截尾与生存时间独立（称独立性截尾）；
- K-M 估计只限于观察生存时间所落的时间区间；

（2）生存率的区间估计

Greenwood 生存率标准误差的近似计算公式为：

$$SE[\hat{S}(t_i)]=\hat{S}(t_i)\sqrt{\sum_{j=1}^{i}\frac{d_j}{n_j(n_j-d_j)}} \tag{15.5}$$

假定生存率近似服从正态分布，则总体生存率的 $(1-\alpha)$ 置信区间为

$\hat{S}(t_i) \pm z_{\alpha/2} \cdot SE[\hat{S}(t_i)]$。

（3）生存率的组间比较

Log rank 检验是生存率比较的非参数方法之一，其基本思想是当 H_0 成立时，根据 t_i 时点的死亡率，可以计算出各组的理论死亡数，则 x^2 统计量计算公式为：

$$\chi^2 = \frac{[\sum w_i(d_{gi} - T_{gi})]^2}{V_g} \tag{15.6}$$

其中，V_g 为第 g 组理论数 T_g 的方差估计，$V_g = \sum w_i^2 \frac{n_{gi}}{n_i}(1-\frac{n_{gi}}{n_i})(\frac{n_i - d_i}{n_i - 1})d_i$。$w_i$ 为权重，对 Log rank 检验，w_i=1。当比较的两总体生存曲线呈比例时，检验效能最大；$w_i = n_i$，则对应 Breslow 检验或 Wilcoxon 检验，该检验给实际死亡率与理论死亡数的早期差别最大的权重。而在 Tarone-Ware 检验中，$w_i = n_i^{1/2}$，其中 n_i 表示时间 t_i 处所对应的期初例数。x^2 近似服从自由度为（组数-1）的 x^2 分布。由于该检验能对各组的生存率做整体比较，因此实际工作中应用较多。

当对多组生存率进行比较时，若分组变量是等级变量，如肿瘤分期为Ⅰ期、Ⅱ期、Ⅲ期，或连续变量的等级化分组，如年龄（岁）<30，30~，40~，≥50，则在 Log rank 检验组间生存率差别有统计学意义后，还可做趋势检验（Tread Test），分析风险率是否有随分组等级变化而变化的趋势。

在 SPSS 中，Kaplan-Meier 分析可由以下步骤实现。

- 选择 Kaplan-Meier 分析对话框

依次单击菜单“ 分析→生存函数→Kaplan-Meier ”命令，打开“ Kaplan-Meier ”对话框，如图 15.7 所示。

- 选择变量

在左侧变量列表框中选择变量后，单击 按钮，可以将其移动到“ 时间 ”、“ 状态 ”、“ 因子 ”或“ 标注个案 ”列表框中，如图 15.7 所示。



图 15.7 “ Kaplan-Meier ”对话框

- “ Kaplan-Meier ”选项

在图 15.7 所示的“ Kaplan-Meier ”对话框中，单击“ 选项 ”按钮，打开“ Kaplan-Meier：选项 ”对话框，如图 15.8 所示。

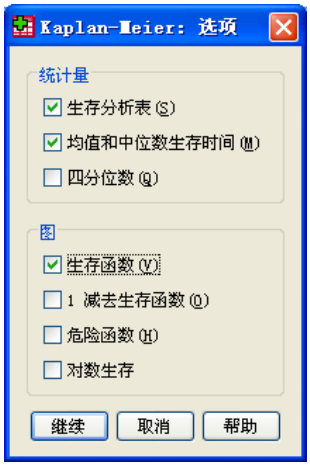


图 15.8 “ Kaplan-Meier：选项 ”对话框

该对话框用于选择需要输出的统计量和统计图，包含两个选项组：“ 统计量 ”和“ 图 ”。在统计选项组中用户可以选择“ 生存分析表 ”、“ 均值和中位数生存时间 ”和“ 四分位数 ”三个复选框。具体含义如表 15.5 所示。

表 15.5 统计量选项及其含义

选　项	含　义
生存分析表	生存率估计表
均值和中位数生存时间	平均生存时间、中位，生存时间（包括标准误差及其置信区间）
四分位数	生存时间的第 25、第 50 和第 75 百分位数

“ 图 ”选项组包括“ 生存函数 ”、“ 1 减去生存函数 ”、“ 危险函数 ”和“ 对数生存 ”四个复选项。其含义如表 15.6 所示。

表 15.6 图选项及其含义

选　项	含　义
生存函数	累积生存函数曲线
1 减去生存函数	累积死亡函数曲线
危险函数	累积风险函数散点图
生存对数	对数累积生存函数曲线

- “ 保存 ”选项

在图 15.7 所示对话框中，单击“ 保存 ”按钮，打开“ Kaplan-Meier：保存新变量 ”对

话框，如图 15.9 所示。

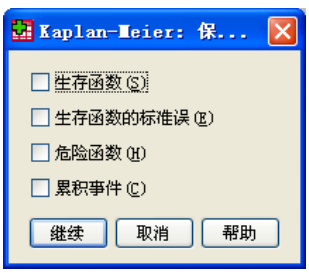
该对话框用于将计算结果保存为新变量，包含“ 生存函数 ”、“ 生存函数的标准误差 ”、“ 危险函数 ” 和 “ 累积事件 ” 四个选项可供用户选择。其含义如表 15.7 所示。

表 15.7　保存选项及其含义

选　项	含　义
生存函数	累积生存函数（生存率）估计值
生存函数的标准误差	累积生存率估计值的标准误差，可用于构造总体生存率的置信区间
危险函数	累积风险率估计
累积事件	累积终点事件发生

- “ 比较因子 ” 选项

在图 15.7 所示对话框中，单击 “ 比较因子 ” 按钮，打开 “ Kaplan-Meier：比较因子水平 ” 对话框，如图 15.10 所示。



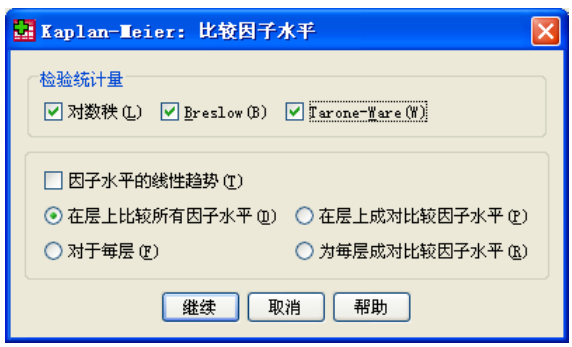


图 15.9 “ Kaplan-Meier：保存新变量 ” 对话框　　图 15.10 “ Kaplan-Meier：比较因子水平 ” 对话框

该对话框用于组间比较，选择具体的统计学检验方法。可供选择的统计量包括 “ 对数秩 ”、“ Breslow ” 和 “ Tarone-Ware ” 三种，它们的主要区别在于观测的权重不同。“ 对数秩 ” 各时间点权重一样，此法最常用；“ Breslow ” 以各时间点的观察例数为权重；“ Tarone-Ware ” 以各时间点观察例数的平方根为权重。

【示例 15.2】Kaplan-Meier 分析示例。

某医生收集 20 例脑瘤患者用甲、乙两种治疗方法的生存时间（周），数据如表 15.8 所示。试估计甲、乙两种治疗方法的生存率并比较两组生存率有无差别。

表 15.8　20 例脑瘤患者两种治疗方法的生存时间（周）

分　组	例　数
甲	5　7^+　13　13　23　30　30^+　38　42　42　45^+
乙	1　3　3　7　10　15　15　23　30

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 整理数据。

将数据整理成如表 15.9 所示格式。

表 15.9　整理后 20 例脑瘤患者两种治疗方法的生存时间（周）

组别	生存时间（time）	生存结局（censor）	组别	生存时间（time）	生存结局（censor）
a	5	1	a	45	0
a	7	0	b	1	1
a	13	1	b	3	1
a	13	1	b	3	1
a	23	1	b	7	1
a	30	1	b	10	1
a	30	0	b	15	1
a	38	1	b	15	1
a	42	1	b	23	1
a	42	1	b	30	1

其中，字符 a、b 分别表示甲、乙两种治疗方法。在“ 生存结局 ”栏中 1 表示死亡，0 表示截尾。

03 Kaplan-Meier 分析。

依次单击菜单“ 分析→生存函数→Kaplan-Meier ”命令，打开“ Kaplan-Meier ”对话框。将“ 生存时间（time）”选入“ 时间 ”列表框；将“ 生存结局（censor）”选入“ 状态 ”列表框；将“ 组别 ”选入“ 因子 ”列表框。

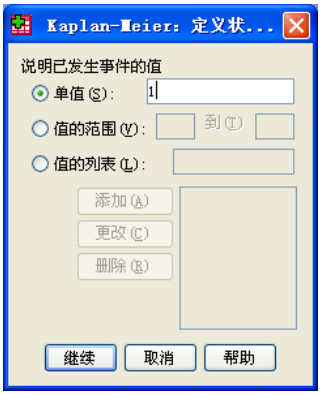


图 15.11　“ Kaplan-Meier：定义状态变量的事件 ”对话框

04 设置。

单击“ 选项 ”按钮，打开“ Kaplan-Meier：选项 ”对话框，在“ 统计量 ”选项组中选择“ 生存分析表 ”和“ 均值和中位数生存时间 ”；在“ 图 ”选项组中选择“ 生存函数 ”。

单击“ 比较因子 ”按钮，打开“ Kaplan-Meier：比较因子水平 ”对话框，在“ 检验统计量 ”选项组中选择“ 对数秩 ”、“ Breslow ”和“ Tarone-Ware ”选项。

单击“ 定义事件 ”按钮，打开“ Kaplan-Meier：定义状态变量的事件 ”对话框，选择“ 单值 ”选项，在文本框中输入 1，如图 15.10 所示。

05 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
KM
生存时间 time  BY 组别 group  /STATUS=生存结局 status(1)
    /PRINT TABLE MEAN .
```

设置完毕后，单击“确定”按钮，执行 Kaplan-Meier 分析操作。

其结果分析如下。

- Case Processing Summary 信息

表 15.10　Case Processing Summary 信息

Case Processing Summary

组别 group	Total N	N of Events	Censored	
			N	Percent
a	11	8	3	27.3%
b	9	9	0	.0%
Overall	20	17	3	15.0%

表 15.10 给出了 a、b 两种治疗方法各组合计的观察例数、死亡数、截尾数及截尾百分比等。

- 两组人群生存率估计表

表 15.11　两组人群生存率估计表

Survival Table

组别 group		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std.Error		
a	1	5.000	1	.909	.087	1	10
	2	7.000	0	.	.	1	9
	3	13.000	1	.	.	2	8
	4	13.000	1	.707	.143	3	7
	5	23.000	1	.606	.154	4	6
	6	30.000	1	.505	.158	5	5
	7	30.000	0		.	5	4
	8	38.000	1	.379	.161	6	3
	9	42.000	1	.	.	7	2
	10	42.000	1	.126	.116	8	1
	11	45.000	0	.	.	8	0
b	1	1.000	1	.889	.105	1	8
	2	3.000	1	.	.	2	7
	3	3.000	1	.667	.157	3	6
	4	7.000	1	.556	.166	4	5
	5	10.000	1	.444	.166	5	4
	6	15.000	1	.	.	6	3
	7	15.000	1	.222	.139	7	2
	8	23.000	1	.111	.105	8	1
	9	30.000	1	.000	.000	9	0

表 15.11 给出了两组人群生存率估计表。其中，Time、Status、Cumulative Survival、Standard Error、Cumulative Event 和 Number Remaining 分别表示生存时间、生存结局、生存率、生存率标准误差、累积死亡数和期初例数。截尾生存时间的生存率和生存率标准误

差与前一个完全生存时间对应值相同，如 a 组 7 周的生存率为 0.909。

- 生存时间均值和中位数信息

表 15.12　生存时间均值和中位数信息

Means and Medians for Survival Time

组别 group	Mean[a]				Median			
	Estimate	Std.Error	95% Confidence Intorval		Estimate	Std.Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
a	29.520	4.352	20.989	38.051	38.000	10.645	17.135	58.865
b	11.889	3.281	5.459	18.319	10.000	4.472	1.235	18.765
Overall	21.347	3.367	14.747	27.947	15.000	5.341	4.532	25.468

a．Estimation is limited to the largest survival time if it is censored.

表 15.12 给出了 a、b 两组平均生存时间、中位生存时间、标准误差及其 95%置信区间。

- 两种治疗方法比较的检验结果

表 15.13　两种治疗方法比较的检验结果

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank（Mantel-Cox）	7.628	1	.006
Breslow（Generalized Wilcoxon）	6.547	1	.011
Tarone-Ware	7.112	1	.008

Test of equality of survival distributions for the different levels of 组别 group.

表 15.13 给出了两种治疗方法比较的检验结果，三种统计检验方法均显示出两组生存率差别有统计学意义。

- 两种治疗方法生存曲线

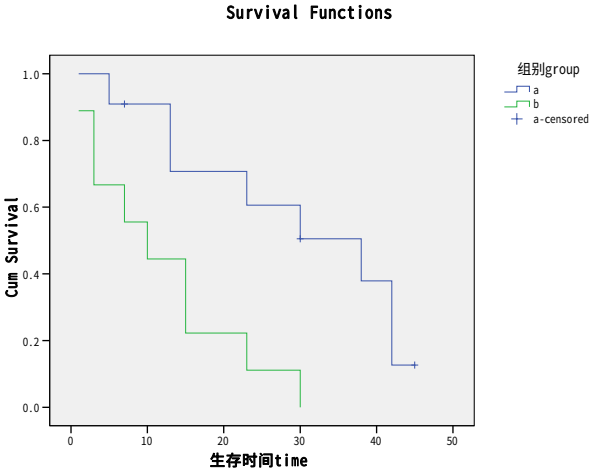


图 15.12　两种治疗方法生存曲线

注释：图 15.12 为两种治疗方法生存曲线图。其中上面的曲线表示 a 组，下面的曲线表示 b 组。从图中可以看出，a 组的治疗方法要优于 b 组的治疗方法。

图 15.12 为两种治疗方法生存曲线图。

15.4　Cox 回归分析

假设有 n 名病人，第 $i(i=1,2,...,n)$ 例病人的生存时间为 t_i，同时设协变量 $X=(X_{i1},X_{i2},...,X_{ip})$ 是影响病人生存时间的 p 个危险因素。设 $h(t,x)$ 表示在受危险因素 x 的影响下，在时刻 t 的风险率；设 $h_0(t)$ 表示在不受危险因素 x 的影响下，在时刻 t 的风险率。显然 $h_0(t)=h(t,0)$，称 $h_0(t)$ 为基础风险系数。

Cox 比例风险模型可写为：

$$h_i(t)=h_0(t)\exp(\sum_{j=1}^{p}\beta_j X_{jj}) \tag{15.7}$$

其中，$h_0(t)$ 是指当所有伴随变量 $X_j(j=1,2,...,p)$ 都处于 0 或标准状态下的风险函数时，为一不确定的值。$\beta_j(j=1,2,...,p)$ 称为 Cox 回归系数，是模型中的待定参数。

对式（15.7）变形后取自然对数有：

$$\ln\frac{h_i(t)}{h_0(t)}=\sum_{j=1}^{p}\beta_j X_{jj} \tag{15.8}$$

如果把式（15.8）的左侧当作因变量，则其形式与一般线性回归类似。因此，人们通常也把 Cox 比例风险模型称为 Cox 回归分析模型。

由式（15.7）可知，变量 x_t 的作用是使个体的风险函数由 $h_0(t)$ 增加到 $h_0(t)\exp(\beta_1)$；P 个协变量 $x_1,x_2,...,x_p$ 共同影响下的风险函数为：

$$h(t,x)=h_0(t)\cdot\exp(\beta_1x_1)\cdot\exp(\beta_2x_2)\cdot...\cdot\exp(\beta_px_p) \tag{15.9}$$

使得个体风险函数由 $h_0(t)$ 增加到 $h_0(t)\cdot\exp(\beta_1x_1)\cdot\exp(\beta_2x_2)\cdot...\cdot\exp(\beta_px_p)$，故 Cox 模型是一种乘法模型。

任何两个个体风险系数之比，即相对危险度 RR 或风险比（Hazard Ratio）可写为：

$$\begin{aligned}RR&=\frac{h_i(t,x)}{h_j(t,x)}=\frac{h_0(t)\exp(\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip})}{h_0(t)\exp(\beta_1x_{j1}+\beta_2x_{j2}+...+\beta_px_{jp})}\\&=\exp[\beta_1(x_{i1}-x_{j1})+\beta_2(x_{i2}-x_{j2})+...+\beta_p(x_{ip}-x_{jp})]\end{aligned} \tag{15.10}$$

其中 $i\neq j,i,j=1,2,...,n$。该比值保持一个恒定的比例，与时间 t 无关，称为比例风险（Proportional Hazards）假定，简称 PH 假定。

由式（15.10）可知，相对风险度 $\frac{h_i(t)}{h_0(t)}$ 的自然对数值 $\ln\frac{h_i(t)}{h_0(t)}$，为伴随变量与对相应回归系数的线性组合。式中 β_j 的实际意义是：当伴随变量 X_j 每改变一个观测单位时，所引起的相对风险度的自然对数值的改变量即为 β_j。例如，在单一自变量情况下，若用 X_j 表示治疗方案，其赋值方式为 $X_{ij}=0$，表示标准治疗方案；$X_{ij}=1$，表示改良治疗方案（i

为病例编号)。这时，一个接受改良治疗方案的病人在时间 t 点的相对风险度的自然对数值 $\ln\frac{h_i(t)}{h_0(t)}$ 为 β_j。显然，当 $\beta_j<0$ 时，有 $h_i(t)<h_0(t)$，这说明改良治疗方案的治疗效果优于标准治疗方案；否则，$h_i(t)>h_0(t)$，即改良治疗方案的治疗效果劣于标准治疗方案。因此，模型中的参数 β_j 不仅反映了伴随变量的作用强度，而且反映了它作用的方向。

同时，式（15.10）也说明在 Cox 比例风险模型中，是假定预后因素对其死亡风险的作用强度在所有时间上都保持一致，这是 Cox 模型的一个重要适用条件。

在 SPSS 中，Cox 回归分析可由以下步骤实现。

- 选择 Cox 回归分析对话框

依次单击菜单“分析→生存函数→Cox 回归”命令，打开“Cox 回归”对话框，如图 15.13 所示。



图 15.13 “Cox 回归”对话框

在对话框下端的“方法”下拉菜单按钮，用户可以通过它来进行方法选择，主要选项及其含义如表 15.14 所示。

表 15.14 方法选项及其含义

方 法 名	含 义
进入	选入“协方差”列表框的全部变量
向前：条件	基于条件参数估计的前进法
向前：LR	基于偏最大似然估计的前进法
向前：Wald	基于 Wald 统计量的前进法
向后：条件	基于条件参数估计的后退法
向后：LR	基于偏最大似然估计的后退法
向后：Wald	基于 Wald 统计量的后退法

- 选择变量

在左侧变量列表框中选择变量后，单击 按钮，可以将其移动到“ 时间 ”、“ 状态 ”、“ 协变量 ” 列表框中。 如图 15.13 所示。

- “ 选项 ” 选项

在图 15.13 所示对话框中，单击“ 选项 ” 按钮，打开“ Cox 回归：选项 ” 对话框，如图 15.14 所示。

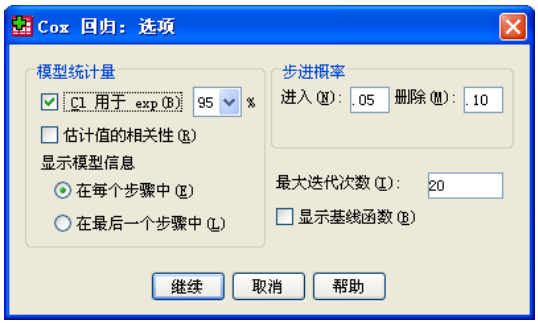


图 15.14 “ Cox 回归：选项 ” 对话框

在该对话框“ 模型统计量 ”选项组中，“ CI 用于 exp(B) ”表示相对危险度的置信区间，系统默认为 95%置信区间；“ 估计值的相关性 ” 表示回归系数的相关矩阵；“ 显示模型信息 ” 有两种输出方式：“ 在每个步骤中 ” 和 “ 在最后一个步骤中 ”。“ 步进概率 ” 表示模型保留变量的显著性水平；“ 显示基线函数 ” 表示输出风险基准函数，以及基于各协变量均值的生存函数与风险函数。

- “ 保存 ” 选项

在图 15.13 所示对话框中，单击“ 保存 ” 按钮，打开“ Cox 回归：保存新变量 ” 对话框，如图 15.15 所示。

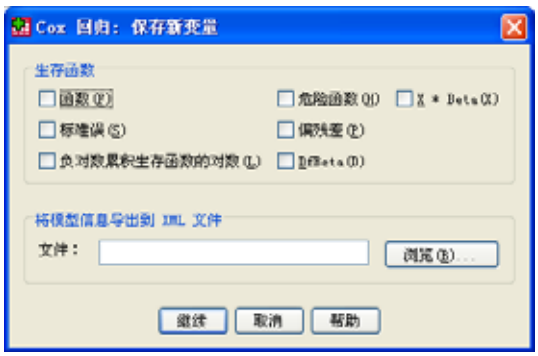


图 15.15 “ Cox 回归：保存新变量 ” 对话框

该对话框用于将计算结果保存为新变量，包含“ 生存函数 ” 和 “ 将模型信息导出到 XML 文件 ” 两个选项可供用户选择。

● “绘图”选项

在图 15.13 所示对话框中，单击“绘图”按钮，打开“Cox 回归：图”对话框，如图 15.16 所示。

该对话框包括四种绘图类型，具体含义如表 15.15 所示。

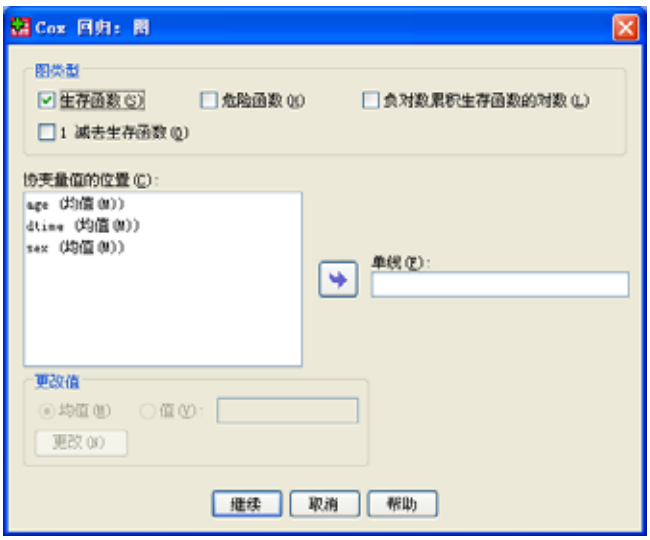


图 15.16 “Cox 回归：图”对话框

表 15.15 图类型及其含义

图 类 型	含 义
生存函数	累积生存函数曲线
危险函数	累积风险函数散点图
负对数累积生存函数的对数	对数累积生存函数乘以-1 后再取对数
1 减去生存函数	生存函数被 1 减后的曲线图

“协变量值的位置”列表框有三个选项，分别是“sex（均值）”、“age（均值）”和“dtime（均值）”。

● “分类”选项

在图 15.13 所示对话框中，单击“分类”按钮，打开“Cox 回归：定义分类协变量”对话框，如图 15.17 所示。

该对话框用于定义分类变量，可将数值变量指定为分类变量，SPSS 自动把它们拆分为 $n-1$ 个变量进行分析（n 为该变量的水平数）。

【示例 15.3】Cox 回归分析示例。

为了了解影响大肠癌患者术后生存情况的因素，30 例手术后的大肠癌患者随访资料如表 15.16 所示。其中术后生存时间 time 以月为单位，status 表示随访结局（0 表示相应的术

后生存时间的缺失值)。3 个协变量分别为：性别 sex（0 表示女性，1 表示男性)、年龄 age（岁)、确诊到进行手术治疗的时间 dtime（月)。试对此数据做 Cox 回归分析。

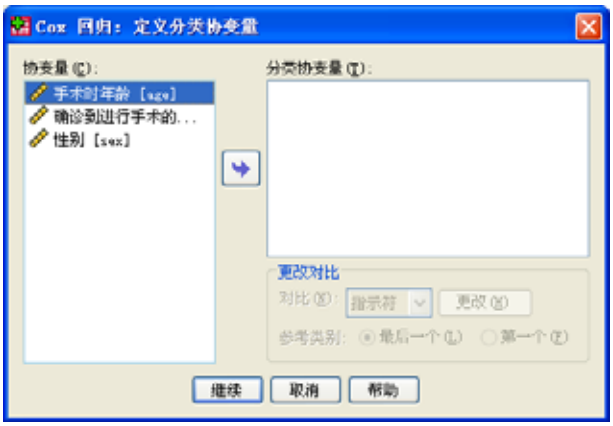


图 15.17 “ Cox 回归：定义分类协变量 ”对话框

表 15.16　30 例大肠癌患者手术后生存资料

time	status	sex	age	dtime	time	status	sex	age	dtime
5	1	0	66	23	8	1	1	66	19
9	1	0	67	21	10	1	1	65	18
12	1	0	63	16	10	1	1	62	22
13	1	0	66	10	12	1	1	64	16
15	1	0	65	15	14	1	1	55	15
16	1	0	59	10	16	1	1	56	8
15	1	0	62	12	19	1	1	58	9
18	1	0	64	9	22	1	1	54	10
20	1	0	58	8	29	1	1	60	7
26	1	0	56	7	32	1	1	55	7
38	1	0	58	10	44	1	1	55	6
41	1	0	53	9	45	1	1	51	8
43	0	0	56	8	56	0	1	5	5
54	1	0	52	6	58	1	1	50	6
59	1	0	48	9	60	0	1	57	3

具体操作步骤如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 Cox 回归分析。

依次单击菜单“ 分析→生存函数→Cox 回归 ”命令，打开“ Cox 回归 ”对话框。将“ 手

术生存时间（time）”选入“ 时间 ”列表框；将“ 随访结局（status）”选入“ 状态 ”列表框；将“ 性别（sex）”、“ 手术时年龄（age）”和“ 确诊到进行手术的时间（dtime）”选入“ 协变量 ”列表框。



图 15.18 “ Cox 回归：为状态变量定义事件 ”对话框

03 设置。

单击“ 选项 ”按钮，打开“ Cox 回归：选项 ”对话框，在“ 模型统计量 ”选项组中选择“ CI 用于 exp(B) ”，置信区间设置为 95%；“ 显示模型信息 ”类型为“ 在每个步骤中 ”。

单击“ 绘图 ”按钮，打开“ Cox 回归：图 ”对话框，在“ 图类型 ”选项组中选择“ 生存函数 ”选项。

单击“ 定义事件 ”按钮，打开“ Cox 回归：为状态变量定义事件 ”对话框，选择“ 单值 ”选项，在文本框中输入 1，如图 15.18 所示。

04 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
COXREG
time /STATUS=status(1)
    /METHOD=FSTEP(LR) sex age dtime
    /PLOT SURVIVAL
    /PRINT=CI(95)
    /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

设置完毕后，单击“ 确定 ”按钮，执行 Cox 回归分析操作。

得到其结果分析如下所示。

- Case Processing Summary 信息

表 15.17 Case Processing Summary 信息

Case Processing Summary

		N	Percent
Cases available In analysis	Event[a]	27	90.0%
	Censored	3	10.0%
	Total	30	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases	0	.0%
	before the earliest event in a stratum	0	.0%
	Total	30	100.0%
Total			

a．Dependent Variable：术后生存时间

注释：表 15.17 为个案处理摘要信息。其中 N 为个案数，Percent 为各种类型个案数所占百分比。Cases availabel in analysis 为在分析中可用的个案数（Eventa 为活动个案数，Censored 为经过检查的个案数），Cases dropped 为忽略个案数（Cases with missing values 为存在缺失值的个案数，Cases with negative time 为拒绝接受的个案数，Censored cases before the earliest event in astratum 为在最早的活动样本中经过检查的个案数）。

表 15.17 给出了总例数、删失例数、失访例数和各自比例等结果。

- 回归方程各参数估计值

表 15.18　回归方程各参数的估计值

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Step 1 dtime	.475	.093	26.007	1	.000	1.608	1.340	1.930
Step 2 age	.234	.068	11.726	1	.001	1.263	1.105	1.444
dtime	.445	.099	20.139	1	.000	1.560	1.285	1.894

表 15.18 是对回归方程各参数的估计，其中 B 为偏回归系数，SE 为偏回归系数的标准误差，Wald 统计量用于检验总体偏回归系数与 0 有无显著性差异。它服从 x^2 分布，当自由度为 1 时，Wald 统计量等于偏自回归系数与标准误差之商的平方。Exp(B)为相对危险度，即 *RR* 值。从表 15.18 给出的逐步回归结果显示，对大肠癌患者生存率有影响的因素是患者年龄和确诊到手术的时间，从回归系数的符合和相对危险度的大小来看，两者都是危险因素。调整确诊到手术时间后，患者年龄每大一岁，手术后死亡的风险将增大到 1.26 倍，即增加 26%；调整年龄后，确诊到手术时间每增加一个月，手术后死亡风险将增加到 1.56 倍，即增加 56%。本例 Cox 模型表达式为：$h(t)=h_0(t)\exp(0.234AGE+0.445ETIME)$。表达式右边指数部分取值越大，则风险系数 $h(t)$ 越大，预后越差，称为预后指数（PI）。此研究提示及早诊断和治疗可延长大肠癌患者的手术后生存期，年轻患者术后效果优于老年患者。

- 未被选入方程的变量

表 15.19　未被选入方程的变量

Variables not in the Equation [a.b]

		Score	df	Sig.
Step 1	sex	.715	1	.398
	Age	5.905	1	.015
Step 2	sex	2.561	1	.110

a．Residual Chi Square = 7.881 with 2 df Sig. = .019

b．Residual Chi Square = 2.561 with 1 df Sig. = .110

表 15.19 显示未被选入方程的变量，按照 Cox 模型的最大似然估计原则，当模型中增加自变量时，L（似然函数值，取值在 0 到 1 之间，其对数 $\ln(L)$ 称为似然对数，取值在负无穷大到 0 之间）将增大，而 $-2\ln(L)$ 将减小，在自变量个数即模型的自由度一定时，$-2\ln(L)$ 取值最小的模型最好，这一点类似于多重线性回归中的剩余平方和。于是我们可以根据模型的 $-2\ln(L)$ 数值大小来考虑自变量的筛选。本例中 3 个自变量都选入不如只选“手术时年龄（age）”和“确诊到进行手术的时间（dtime）”这两个变量建立模型，所以变量“性别（sex）”未被选入。

- 各协/变量的均值

表 15.20　各协/变量的均值

Covariate Means

	Mean
sex	.500
age	56.867
dtime	11.067

表 15.20 所示的是各协/变量的均值。

- 基于各协变量均值的生存曲线

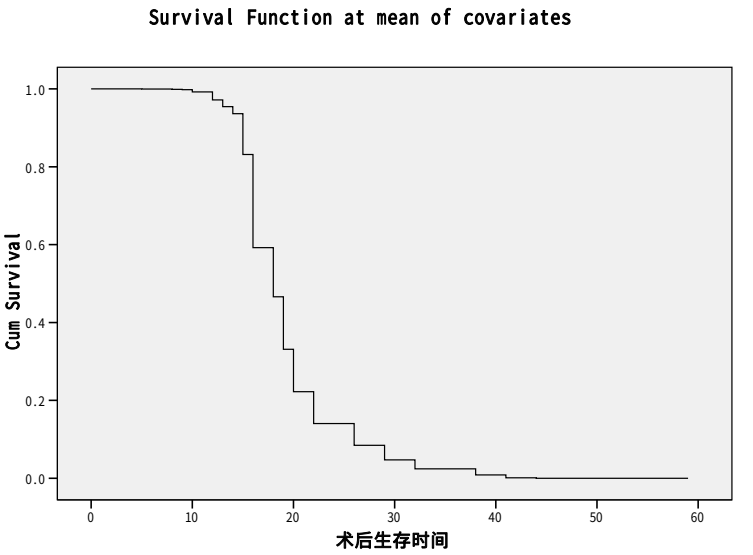


图 15.19　基于各协/变量均值的生存曲线

图 15.19 是在各协变量均值水平时的累积生存函数曲线，其意义在于研究样本所在总体人群总的生存率变化情况。本例大肠癌术后 30 个月以上的生存率非常低。

15.5　本章小结

本章内容为生存分析，包括生存表分析、Kaplan-Meier 分析和 Cox 回归分析。生存分析是专业统计分析中最重要的统计分析方法，已经得到了广泛的应用。

本章所述的 SPSS 操作比较简单，读者应重点掌握生存分析的基本思想以及统计结果输出的实际含义。

第三篇

SPSS 17.0 行业应用实例

SPSS 广泛应用于通讯、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，是世界上应用最广泛的专业统计软件。本篇将通过 10 个典型案例详细介绍 SPSS 在工程中的应用，具体如下。

- 第 16 章　SPSS 行业应用实例——社会调查与统计
- 第 17 章　SPSS 行业应用实例——市场研究/企业数据分析
- 第 18 章　SPSS 行业应用实例——证券金融统计
- 第 19 章　SPSS 行业应用实例——医学统计
- 第 20 章　SPSS 行业应用实例——生物学统计

第16章 SPSS行业应用实例——社会调查与统计

SPSS 是应用非常普遍的社会统计软件，它具有专业级的统计分析功能和强有力的数据管理能力，同时又具有易用易学的特点，在社会调查与统计中发挥着重要的作用。SPSS 应用于社会调查与统计，能使研究者以客观的态度，通过对被调查人的系统提问，收集并分析有关研究数据，以描述、解释或预测问卷调查内容的现象及其各相关因素之间的关系，从而减少从事社会科学研究、经济管理、商业决策、人文教育等行业的中高层管理人员的劳动量，提高工作效率，也可以使分析结果更具有科学性，为企业或政府提供决策支持。

16.1 全国各地区农民家庭收支的分布规律分析

16.1.1 实例内容说明

研究某年全国各地区农民家庭收支的分布规律，根据抽样调查资料进行分类处理，抽取 28 个省、市、自治区的样品，每个样品有 6 个指标。先采用 SPSS 系统聚类分析将 28 个样本分为 3 类，其中有 3 个样本（北京、上海、广州）属于孤立样本，未归属于已分的 3 类中，现要求对这 28 个样本的所属类别进行分析。原始数据如表 16.1 所示。

表 16.1 28 个地区农民家庭收支数据表

类别	序号	地区	食品	衣着	燃料	住房	生活用品及其他	文化生活服务支出
一组	1	天津	36.40	10.47	44.16	36.40	3.94	36.40
	2	辽宁	32.83	17.79	27.29	39.09	3.47	32.83
	3	吉林	33.38	18.37	11.81	25.29	5.22	33.38

续表

类别	序号	地区	食品	衣着	燃料	住房	生活用品及其他	文化生活服务支出
	4	江苏	29.12	11.67	42.60	27.30	5.74	29.12
	5	浙江	32.75	12.72	47.12	34.35	5.00	32.75
	6	山东	30.76	12.20	33.61	33.77	3.85	30.76
二组	7	黑龙江	29.57	13.24	13.76	21.75	6.04	29.57
	8	安徽	23.09	15.62	23.54	18.18	6.39	23.09
	9	福建	21.06	16.96	19.52	21.75	6.73	21.06
	10	江西	21.59	17.64	19.19	15.97	4.94	21.59
	11	湖北	28.26	12.35	18.53	21.95	6.23	28.26
	12	湖南	24.74	13.63	22.20	18.06	6.04	24.74
	13	广西	18.47	14.68	13.41	20.66	3.85	18.47
	14	四川	20.74	11.07	17.74	16.49	4.39	20.74
	15	贵州	21.53	12.58	14.49	12.18	4.57	21.53
	16	新疆	38.00	13.72	4.64	17.77	5.75	38.00
三组	17	河北	22.83	9.30	22.44	22.81	2.80	22.83
	18	山西	25.11	6.46	9.89	18.17	3.25	25.11
	19	内蒙	27.63	8.94	12.58	23.99	3.27	27.63
	20	河南	23.26	8.46	20.20	20.50	4.30	23.26
	21	云南	19.81	8.89	14.22	15.53	3.03	19.81
	22	陕西	20.56	10.94	10.11	18.00	3.29	20.56
	23	甘肃	16.82	5.70	6.03	12.36	4.49	16.82
	24	青海	16.45	8.98	5.40	8.78	5.93	16.45
	25	宁夏	24.11	6.46	9.61	22.92	2.53	24.11
待判别样品	26	北京	43.77	9.73	60.54	49.10	9.04	43.77
	27	上海	38.64	12.53	115.65	50.82	5.89	38.64
	28	广州	20.52	18.32	42.40	36.97	11.68	20.52

16.1.2　实现方法分析

要对表 16.1 中 28 个样本数据类别进行分析，可以采用聚类分析和判别分析。聚类分析又称群分析，是分类学的一种基本方法。通俗地讲，“ 类 ”就是由具有相似性的元素构成的集合，聚类分析也是多元统计学中应用极为广泛的一种重要方法。判别分析（Discriminant Analysis）是类别明确的一种分类技术，它根据观测到的某些指标对所研究的对象进行分类，得到所谓的判别函数，然后在使用判别函数对未知分类的样品进行分类。

聚类分析是在不知道类别数目的情况下对样本数据进行分类，而判别分析则是在已知分类数据的情况下，根据一定的指标对不知类别的数据进行归类。从题目中可以看出，

28 个样本中有 25 个的类别已经确定，北京、上海和广州这三个样本属于未知类别样本，现在需要对它们进行分析，所以本例选用判别分析方法来判定这 28 个样品的所属类别。

16.1.3 具体操作步骤

根据前面分析，本例需要用判别分析对这 28 个样本进行类别归属分析。

16.1.3.1 双变量相关分析

在 SPSS 中进行判别分析，可以利用“ 分析→分类→判别 ”命令实现，菜单内容如图 16.1 所示。

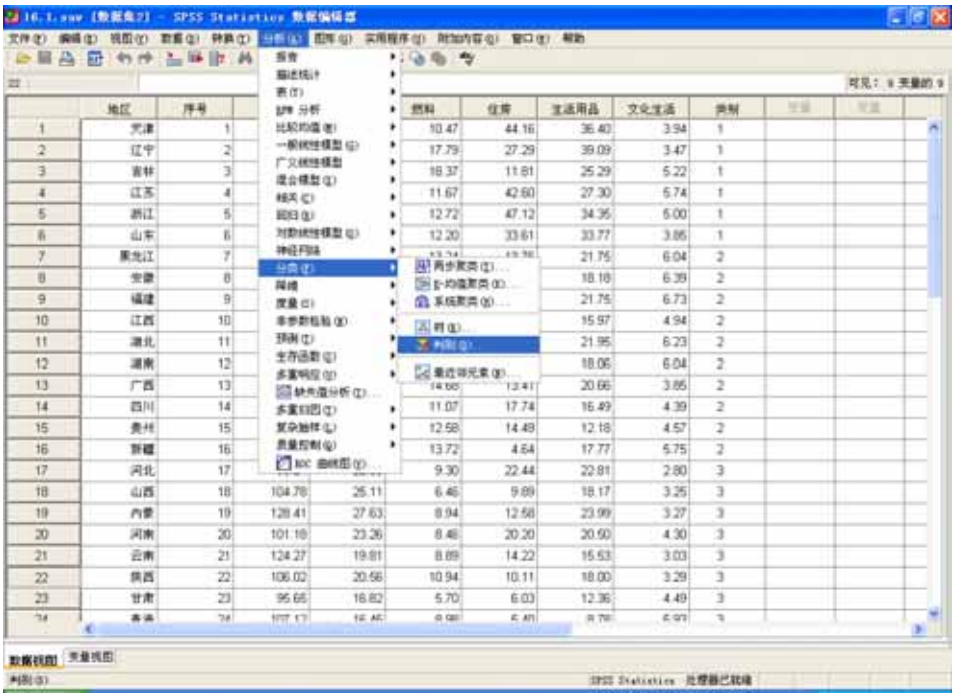


图 16.1 SPSS 判别分析模块

判别分析的 SPSS 操作如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“ 分析→分类→判别 ”命令，打开“ 判别分析 ”对话框。如图 16.2 所示。

在对话框的左侧变量列表框中单击选中变量“ 序号 ”，单击按钮，将其移到右侧的“ 分组变量 ”列表框；将变量“ 食品[食品] ”、“ 衣着[衣着] ”、“ 燃料[燃料] ”、“ 住房[住房] ”、“ 生活用品[生活用品] ”和“ 文化生活[文化用品] ”选入“ 自变量 ”列表框，如图 16.3 所示。



图 16.2　判别分析对话框



图 16.3　选入变量后对话框

单击图 16.3 中“ 定义范围 ”按钮，打开“ 判别分析：定义范围 ”对话框，将“ 最小值 ”设为 1，“ 最大 ”设为 3。如图 16.4 所示。

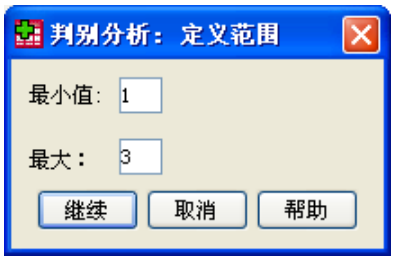


图 16.4　判别分析：定义范围对话框

在图 16.2 所示对话框中可以指定判别分析的方法，其内容及意义如表 16.2 所示。

表 16.2　判别分析方法选项及意义

方　法	意　义
一起输入自变量	默认选项。当认为所有自变量都能对观测量特性提供丰富的信息时，使用该选项，选择该项将不加选择地使用所有自变量进行判别分析，建立全模型，且不需要进一步选择
使用步进式方法	当认为不是所有自变量都能对观测量特性提供丰富的信息时，选择该项，因此需要判别贡献的大小再进行选择。选中该单选按钮后，“ 方法 ”按钮将被激活，可以进一步选择判别分析方法

本例选择“ 使用步进式方法 ”。

- 判别分析：统计量对话框

单击图 16.2 中“ 统计量 ”按钮，打开“ 判别分析：统计量 ”对话框。如图 16.5 所示。

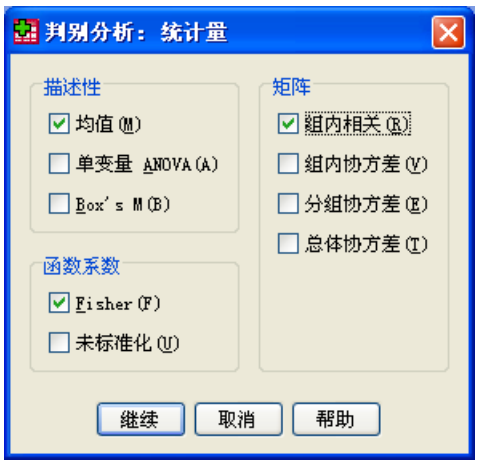


图 16.5 “ 判别分析：统计量 ”对话框

该对话框给出了输出结果中显示的统计量，包括如下选项：

在“ 描述性 ”选项组中选择对原始数据的描述统计量的输出，其选项及含义如表 16.3 所示。

表 16.3 “ 描述性 ”选项及含义

选　项	含　义
均值	选择该项，可以输出各类中各自变量的均值、标准差以及各自变量总样本的均值和标准差
单变量 ANOVA	选择该项，表示对每一类同一自变量均值都相等的假设进行检验，输出单因变量的方差分析结果
Box's M	选择该项，表示对各类的协方差矩阵相等的假设进行检验

“ 函数系数 ”选项组中给出选择判别函数系数的输出形式的选项，其选项及含义如表 16.4 所示。

表 16.4 “ 函数系数 ” 选项及含义

选　项	含　义
Fisher	选择该项，表示可以直接用于对新样本进行判别分类的 Fisher 系数，对每一类给出一组系数，并给出该组中判别分数最大的观测量
未标准化	选择该项，表示未经标准化处理的判别系数

在 “ 矩阵 ” 选项组中选择自变量的系数矩阵，有 4 个复选项。如表 16.5 所示。

表 16.5 “ 矩阵 ” 选项及含义

选　项	含　义
组内相关	在计算相关矩阵之前将个案协方差矩阵平均后，计算类内相关矩阵
组内协方差	将个案协方差矩阵平均后计算的，区别于总协方差矩阵
分组协方差	协方差矩阵
总体协方差	总样本的协方差矩阵

本例中 “ 描述性 ” 选项组选择 “ 平均值 ”；“ 矩阵 ” 选项组选择 “ 组内相关 ”；“ 函数系数 ” 选项组选择 “ Fisher ” 选项。

- 判别分析：步进法对话框

单击图 16.2 中 “ 方法 ” 按钮，打开 “ 判别分析：步进法 ” 对话框。如图 16.6 所示。



图 16.6 “ 判别分析：步进法 ” 对话框

在图 16.6 所示对话框中，“ 方法 ” 选项组用于选择逐步判别分析的方法，可供选择的判别分析方法有 5 种，如表 16.6 所示。

表 16.6 “ 方法 ” 选项及含义

方　法	含　义
Wilks’lambda	默认选项，每步都是 Wilk 的概率量最小的进入判别函数
未解释方差	选择该项，表示每步都是使各类不可解释的方差和最小的变量进入判别函数
Mahalanobis 距离	选择该项，表示每步都使靠得最近的两类间的 Mahalanobis 距离最大的变量进入判别函数

续表

方　法	含　义
最小值	选择该项，表示每步都使任何两类间的最小的 *F* 值最大的变量进入判别函数
Rao'sV	选择该项，表示每步都使 Rao *V* 统计量产生最大增量的变量进入判别函数，需要在该项下面的“ *V* 值输入 ”文本框中输入这个增量的指定值，当某变量导致的 *V* 值增量大于指定值的变量后进入判别函数

“ 标准 ”选项组用于选择逐步判别停止的依据，可供选择的选项及含义如表 16.7 所示。

表 16.7 “ 标准 ” 选项及含义

方　法	意　义
使用 F 值	默认选项。使用 *F* 值是系统默认的依据，当加入一个变量（或剔除一个变量）后，对判别函数中的变量进行方差分析。当计算的 *F* 值大于指定的“ 进入 ”值时，该变量保留在函数中，默认“ 进入 ”的值是 3.84；当该变量使计算的 *F* 值小于指定的移除值时，该变量从函数中剔除，默认“ 移除 ”值为 2.71。即当被加入的变量 *F* 的值为 3.84 的时，才把该变量加入到模型中，否则变量不能进入模型；或者当要从模型中移除的变量的 *F* 的值为 2.71 时，该变量才被移出模型，否则模型中的变量不会被移出。设置这两个值时应该注意“ 进入 ”值和“ 移除 ”值
使用 F 的概率	选择该项，表示用 *F* 检验的概率决定变量是否被加入函数或被剔除，而不是用 *F* 值加入变量的，*F* 值概率的默认值是 0.05，移出变量的 *F* 值概率是 0.10。“ 移除 ”值是移出变量的 *F* 值概率；“ 进入 ”值是加入变量的 *F* 值概率

“ 输出 ”选项组的选项用于显示逐步变量判别法的过程设置，可供选择的如表 16.8 所示的两个复选项。

表 16.8 “ 输出 ” 选项及含义

显示选项	含　义
步进摘要	要求在逐步选择变量过程中的每一步之后显示两个变量的统计量
两两组间距离的 F 值	要求显示两类之间的两两 *F* 值矩阵

本例中“ 方法 ”选项组选择“ Wilks' lambda ”方法；“ 标准 ”选项组选择“ 使用 F 分布值 ”单选项，“ 进入 ”设置为 2.5，“ 移除 ”设置为 2.49；“ 输出 ”选项组选择“ 步进摘要 ”。

- 判别分析：分类对话框

单击图 16.2 中所示“ 分类 ”按钮，打开“ 判别分析：分类 ”对话框。如图 16.7 所示。

“ 判别分析：分类 ”对话框用于指定分类参数和判别结果。下面分别介绍各个选项组的选项。

在“ 先验概率 ”选项组中选择先验概率，有两个单选项供选择，如表 16.9 所示。



图 16.7　判别分析：分类对话框

表 16.9 “ 先验概率 ” 选项及含义

选　　项	含　　义
所有组相等	表示个案先验概率相等，若分为 *m* 类，则各类先验概率均为 *1/m*
根据组大小计算	表示由各类的样本量计算决定，即各类的先验概率与其样本量成正比

在 “ 使用协方差矩阵 ” 选项组中选择分类使用的协方差矩阵，有表 16.10 所示两个单选项。

表 16.10 “ 使用协方差矩阵 ” 选项及含义

选　　项	含　　义
在组内	选择该项，表示指定使用合并组内协方差矩阵进行分类
分组	选择该项，表示指定使用各组协方差矩阵进行分类，由于分类是根据判别函数而不是根据原始变量，因此该选项不是总等价于二次判别

在 “ 图 ” 选项组中选择要求输出的统计图形，有表 16.11 所示三个单选项。

表 16.11 “ 图 ” 选项及含义

选　　项	含　　义
合并组	选择该项，生成一张包含各类的散点图，该散点图是根据前两个判别函数值做的散点图；如果只有一个判别函数，就输出直方图
分组	选择该项，根据前两个判别函数值对每一类生成一张散点图，共分为几类就生成几张散点图；如果只有一个判别函数，就输出直方图
区域图	选择该项，生成用于根据函数值把观测量分到各组中去的边界图，此种统计图把一张图的平面划分出与类数相同的区域，每一类占据一个区，各类的均值在各区中用*号标出；如果仅有一个判别函数则不作此图

在 “ 输出 ” 选项组中选择要求输出的统计图形，有表 16.12 所示三个单选项。

表 16.12 “ 输出 ” 选项及含义

选　　项	含　　义
个案结果	要求输出每个观测量，包括判别分数实际类预测类（根据判别函数求得的分类结果）和后验概率等，选择此项还可以选择其附属选择项，选择“ 限制观测量到最初 ” 选项，并在后面的文本框中输入观测量数 *n*，选择此项仅对前 *n* 个观测量输出分类结果，观测数量大时可以选择此项
摘要表	要求输出分类的综述表，给出正确分类观测量数（原始类和根据判别函数计算的预测类相同）和错分观测量数及错分率
不考虑该个案时的分类	输出对每个观测量进行分类的结果，所依据的判别是由除该观测量以外的其他测量导出的，也称为交互检验结果

本例中“ 先验概率 ” 选项组选择“ 所有组相等 ”；“ 使用协方差矩阵 ” 选项组选择“ 在组内 ”；“ 输出 ” 选项组选择“ 个案结果 ”，不设定“ 将个案限制在前 ”；“ 图 ” 选项组选择“ 合并组 ” 和“ 地域图 ”。

- 判别分析：保存对话框

单击图 16.2 中“ 保存 ” 按钮，打开“ 判别分析：保存 ” 对话框。如图 16.8 所示。



图 16.8　判别分析：保存对话框

在“ 判别分析：保存 ” 对话框中，用户可以选择要保存的变量和选择保存的路径。其内容如表 16.13 所示。

表 16.13 “ 保存 ” 选项及含义

选　　项	含　　义
预测组成员	选择该项，要求建立一个新变量预测观测量的分裂，是根据判别分数把观测量按后验概率最大指派所属的类，每运行一次判别过程就建立一个，表明使用判别函数预测各观测量属于哪一类的新变量。第 1 次运行建立新变量的变量名为 dis_1，如果在工作数据文件中不把前一次建立的新变量删除，第 *n* 次运行判别过程建立的新变量默认的变量名为 dis_*n*
判别得分	选择该项，要求建立表明判别分数的新变量，该分数是由未标准化的判别系数和自变量的乘积加上常数得来的。每次运行判别过程都给出一组表明判别分数的新变量，建立几个判别函数就有几个判别分数变量参与分析的观测量，共分为 *m* 类，则建立 *m* 个点，判别函数指定该选择项就可以生成 *m*–*1* 个表明判别分数的新变量
组成员概率	选择该项，要求建立新变量表明观测量属于某一类的概率。如果有 *m* 类，对一个观测量就会给出 *m* 个概率值，因此建立 *m* 个新变量

本例将其全部选中。

03 单击“粘贴”按钮,得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 1.
DISCRIMINANT
  /GROUPS=序号(1 3)
  /VARIABLES=食品 衣着 燃料 住房 生活用品 文化生活
  /ANALYSIS ALL
  /SAVE=CLASS SCORES PROBS
  /METHOD=WILKS
  /FIN=2.5
  /FOUT=2.49
  /PRIORS EQUAL
  /HISTORY
  /STATISTICS=MEAN STDDEV COEFF CORR
  /PLOT=COMBINED MAP
  /PLOT=CASES
  /CLASSIFY=NONMISSING POOLED.
```

04 结果输出。

设置完毕后，单击“确定”按钮，执行判别分析。

- 个案处理摘要

表 16.14　个案处理摘要

分析案例处理摘要

未加权案例		N	百分比
有效		25	89.3
排除的	缺失或越界组代码	3	10.7
	至少一个缺失判别变量	0	.0
	缺失或越界组代码还有至少一个缺失判别变量	0	.0
合计		3	10.7
合计		28	100.0

表 16.14 为参与分析的个案处理摘要表，表中列出了有效值（Valid）、缺失值（Exclude）总和（Total）的个数（N）和百分比（Percent）。

- 特征值

表 16.15　特征值

特征值

函　数	特 征 值	方差的%	累积%	正则相关性
1	5.345[a]	81.8	81.8	.918
2	1.193[a]	18.2	100.0	.738

a．分析中使用了前 2 个典型判别函数。

表 17.15 为特征值，第二列至第四列分别表示各项为前两个判别函数的特征值占总方差的百分数、累加百分比和正规相关系数。最大特征值对应于组均值最大扩展方向上的特征值向量，第二大特征值对应于组均值次大扩展方向上的特征值向量，依次类推。特征值的平方根提供了对应特征值向量的长度信息，或该维的典则变量均值的宽度。这里，第一个典则变量解释了 81.8%的总方差。

- Wilks 的 λ 值表

表 16.16　Wileks 的 λ 值表

Wilks 的 Lambda

函数检验	Wioks 的 Lambda	卡　方	df	Siq.
1 到 2	.072	53.976	8	.000
2	.456	16.098	3	.001

表 16.16 为 Wilk λ 值表。表中各列分别是 Wilk λ 值、卡方值、自由度和显著性概率。当显著性概率小于 0.05 时，拒绝原假设，认为组间均值不相等。表中第一行假设 3 个组中两个函数的均值相等，第二行则在剔除了第一个函数以后检验 3 个组中函数的均值相等。

- 标准化的典型判别式函数系数

表 16.17　标准化的典型判别式函数系数

标准化的典型判别式函数系数

	函　数	
	1	2
衣着	.640	.2003
燃料	.842	.676
住房	.857	.166
生活用品	–.093	–.941

表 16.17 为标准化的典型判别式函数系数，包括各独立变量对应的判别函数的标准化的系数值。根据该表可以得到判别函数如下：

$$f = 0.640\times 衣着 + 0.842\times 燃料 + 0.857\times 住房 - 0.093\times 生活用品$$

$$f = 0.203\times 衣着 + 0.676\times 燃料 + 0.166\times 住房 - 0.941\times 生活用品$$

注意，判别函数的个数为 $k-1$ 或 p 中的最小值，式中 k 为组数， p 为变量个数。这里，由于组数为 3，所以判别函数的个数为 2，它比变量个数小。

- 结构矩阵

表 16.18　结构矩阵

结构矩阵

	函　数	
	1	2
住房	.453*	–.437
衣着	.397*	–.301
食品 a	.334*	.197
生活用品	.511	–.784*
燃料	.482	.587*
文化生活 a	.052	.431*

判别变量和标准化典型判别式函数之间的汇聚组之间相关性按函数内相关性的绝对大小排序的变量。

*．每个变量和任意判别式函数间最大的绝对相关性。

a．该变量不在分析中使用。

表 16.18 为结构矩阵。表内为判别变量与根据函数内相关系数绝对值大小排序的标准化公共判别函数变量之间的合并组内相关系数。

- 组质心处的函数表

表 16.19　组质心处的函数表

组质心处的函数

序　号	函　数	
	1	2
1	3.068	–1.106
2	.463	1.236
3	–.2.560	–.636

在组均值处评估的非标准化典型判别式函数。

表 16.19 为组质心处的函数值表，包括对应于前两个函数的各类别的函数值。

- 组统计量表

表 16.20　组统计量表

组统计量

序　号		均　值	标 准 差	有效的 N（列表状态）	
				未加权的	已加权的
1	食品	145.1650	18.83692	6	6.000
	衣着	32.5400	2.47379	6	6.000
	燃料	13.8700	3.35046	6	6.000
	住房	34.4317	13.30995	6	6.000

续表

序　号		均　值	标 准 差	有效的 N（列表状态）	
				未加权的	已加权的
	生活用品	32.7000	5.33774	6	6.000
	文化生活	4.5387	.90498	6	6.000
2	食品	138.1240	14.63105	10	10.000
	衣着	24.7050	5.80554	10	10.000
	燃料	14.1490	2.08147	10	10.000
	住房	16.7020	5.43051	10	10.000
	生活用品	18.4760	3.14289	10	10.000
	文化生活	5.4930	.97845	10	10.000
3	食品	108.4867	11.67144	9	9.000
	衣着	21.8422	3.74797	9	9.000
	燃料	8.2367	7.68487	9	9.000
	住房	12.2756	5.84970	9	9.000
	生活用品	18.1178	5.15140	9	9.000
	文化生活	3.6544	1.06780	9	9.000
合计	食品	129.1444	21.38682	25	25.000
	衣着	25.5548	6.02199	25	25.000
	燃料	11.9536	3.60789	25	25.000
	住房	19.3636	11.74419	25	25.000
	生活用品	21.7608	7.60754	25	25.000
	文化生活	4.6016	1.25628	25	25.000

表 16.20 为组统计量表。表中列出了各个类别以及总和对应的各个变量的均值（Mean）、标准偏差（Std.Deviation）和未加权（Weighted）的有效个案数（Valid N）。

在“判别分析：统计量”和“判别分析：分类”对话框中，选中所有复选框，单击“确定”按钮，得到如下运行结果。

- 组均值的均等性检验表

表 16.21　组均值的均等性检验表

组均值的均等性的

	Wilks 的 Lambda	F	dC	dr2	8ia.
食品	.436	14.206	2	22	.000
衣着	.513	10.451	2	22	.001
燃料	.377	18.P64	2	22	.000
住房	.430	14.554	2	22	.000
生活用品	.3P9	23.439	2	22	.000
文化生活	.576	8.084	2	22	.002

表 16.21 为组间均值相等性检验表。包括变量对应的 Wilk λ 值、F 值、第一自由度（df1）、第二自由度（df2）和显著性概率（Sig.）。该检验的零假设的组均值相等，没有显著差异。由于表 16.21 中各变量对应的显著性概率均小于 0.05，所以拒绝原假设，认为各变量的组均值存在显著差异。

- 汇聚的组内矩阵

表 16.22　汇聚的组内矩阵

汇集的组内矩阵

		食品	衣着	燃料	住房	生活用品	文化生活
协方差	食品	217.752	–6.242	10.758	18.176	1.587	2.407
	衣着	–.6242	20.287	–1.117	–4.749	8.820	–.162
	燃料	10.758	–1.117	5.356	–6.329	.112	.153
	住房	18.176	–4.749	–6.329	64.770	13.334	–.191
	生活用品	1.587	8.820	.112	13.334	20.166	–1.890
	文化生活	2.407	–.162	.153	–.191	–1.890	.992
相关性	食品	1.000	–.094	.315	.153	.024	.164
	衣着	–.094	1.000	–.107	–.131	.436	–.036
	燃料	.375	–.107	1.000	–.340	.011	.066
	住房	.753	–.131	–.340	1.000	.369	–.024
	生活用品	.024	.436	.011	.369	1.000	–.422
	文化生活	.P64	–.036	.066	–.024	–.422	1.000

a．协方差矩阵的自由度为 22。

表 16.22 为汇聚的组内矩阵，表中为各变量两两之间的协方差和相关系数。

- 协方差矩阵

表 16.23　协方差矩阵

协方差矩阵[a]

序　号		食品	衣着	燃料	住房	生活用品	文化生活
1	食品	354.830	6.329	24.147	–17.056	–25.674	9.297
	衣着	6.329	6.120	.410	–1.466	5.364	–.967
	燃料	24.147	.410	11.226	–38.812	–2.536	–.222
	住房	–17.056	–1.466	–38.812	177.155	26.787	.525
	生活用品	–25.674	5.364	–2.536	26.787	28.492	–4.111
	文化生活	9.297	–.967	–.222	.525	–4.111	.819
2	食品	214.068	–33.725	9.470	61.277	6.094	3.855
	衣着	–33.725	33.704	–2.906	–19.291	3.275	2.448
	燃料	9.470	–2.906	4.333	2.989	.941	.513
	住房	61.277	–19.291	2.989	29.490	.954	1.577
	生活用品	6.094	3.275	.941	.954	9.878	1.503
	文化生活	3.855	2.448	.513	1.577	1.503	.957

续表

序　　号		食品	衣着	燃料	住房	生活用品	文化生活
2	食品	136.222	16.821	3.838	–8.293	13.556	–3.529
	衣着	16.821	14.047	–.057	9.558	17.217	–2.594
	燃料	3.836	–.057	2.839	3.489	.835	–.017
	住房	–8.293	9.558	3.489	34.219	18.853	–2.627
	生活用品	13.556	17.217	.835	18.853	26.537	–4.319
	文化生活	–3.529	–2.594	–.017	–2.627	–4.319	1.140
合计	食品	457.396	47.836	54.545	121.961	61.199	12.619
	衣著	47.836	36.264	6.720	32.770	33.423	.741
	燃料	54.545	6.720	13.017	8.862	7.417	2.245
	住房	121.961	32.770	8.862	137.926	66,757	1.110
	生活用品	61.199	33.423	7.417	66.757	57.875	–1.836
	文化生活	12.619	.741	2.245	1.110	–1.836	1.570

a．总的协方差矩阵的自由度为 24。

表 16.23 为协方差矩阵，包括各个类别和全部对应的各变量两两之间的协方差。

- 对数行列式

表 16.24　对数行列式

对数行列式

序　　号	秩	对数列行式
1	4	10.725
2	4	9.998
3	4	8.115
汇聚的组内	4	11.140

打印的行列式的秩和自然对数是组协方关矩阵的秩和自然对数。

表 16.24 为对数行列式表，包括各类别和合并组内对应的秩和对数行列式。

- Box’s M 检验成果表

表 16.25　协方差矩阵相等性检验——Box’s M 检验成果表

检验结果

箱的 M	36.557
F　近似。	1.282
df1	20
df2	1074.568
Sig.	.181

检验结果总体协方差矩阵的零假设进行检验。

表 16.25 为协方差矩阵相等性检验——Box's M 检验成果表。表中类出了 Box's M 值、近似 F 值、第一自由度、第二自由度和显著性概率。零假设为组的协方差矩阵相等。由于显著性概率 ρ 值为 0.181，大于 0.05，所以接受原假设，认为协方差矩阵相等，即差异不显著。

- 典型判别式函数系数

表 16.26　典型判别式函数系数

典型判别式函数系数

	函　数	
	1	2
衣着	.142	.045
燃料	.364	.292
住房	.106	.021
生活用品	–.021	–.209
(常量)	–9.593	–.483

非标准化系数

表 16.26 为典型判别函数系数。表中列出的是各变量及常数项对应的没有进行标准化的典型判别函数系数。利用该表，可以得到前两个判别函数的表达式如下：

$$f = 0.142*\text{衣着} + 0.364*\text{燃料} + 0.106*\text{住房} - 0.021*\text{生活用品} - 9.593$$

$$f = 0.045*\text{衣着} + 0.292*\text{燃料} + 0.021*\text{住房} - 0.209*\text{生活用品} - 1.429$$

- 组的先验概率

表 16.27　组的先验概率

组的先验概率

序　号	先　验	用于分析的窦例	
		未加权的	已加权的
1	.333	6	6.000
2	.333	10	10.000
3	.333	9	9.000
合计	1.000	25	25.000

表 16.27 为组的先验概率。包括各类别和全部对应的先验概率、参与分析的未加权和经过加权的个案数。

- Fisher's 线性判别函数系数

表 16.28 为 Fisher's 线性判别函数系数表。利用该表得到 3 个类别的分类判别函数为：

f_1=2.154*衣着+4.386*燃料+1.138*住房–0.098*生活用品–84.567

f_2=1.890*衣着+4.121*燃料+0.909*住房–0.534*生活用品–56.255

f_3=1.376*衣着+2.475*燃料+0.549*住房–0.080*生活用品–28.964

表 16.28　Fisher's 线性判别函数系数

分类函数系数

	序　号		
	1	2	3
衣着	2.P54	1.890	1.376
燃料	4.386	4.121	2.475
住房	1.138	.909	.549
生活用品	−.098	−.534	−.080
(常量)	−84.567	−56.255	−28.964

Fisher 的线性性判别式函数

将 3 个待判样本的变量数据代入 3 个函数中，对每个样本对应的 3 个函数的值进行比较，第几个函数的值最大，则该样本判为第几类。计算后会发现北京、上海和广州 3 个城市对应的 3 个函数值都是第 1 个函数值大，所以它们划分为第一类。

- 区域图

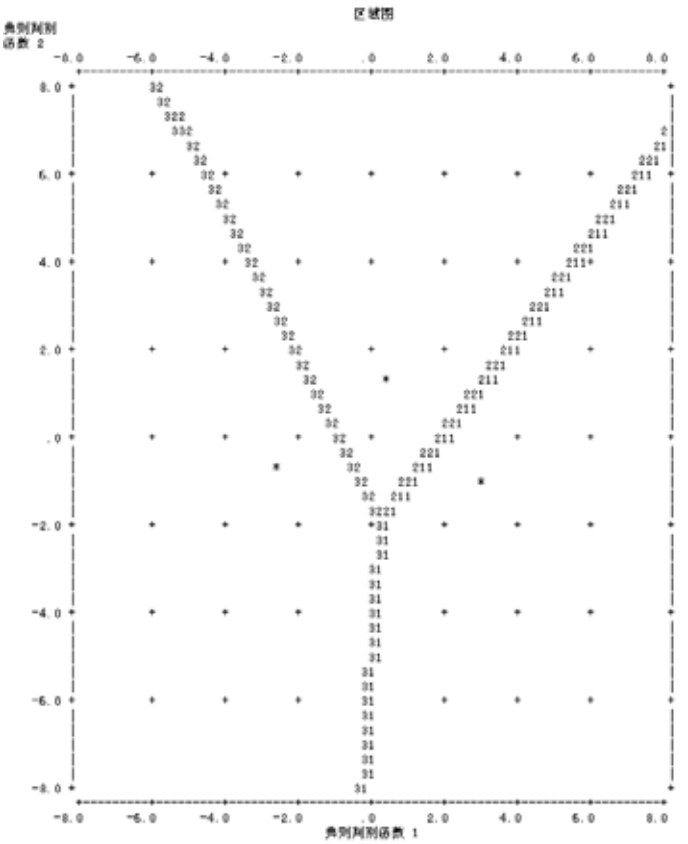


图 16.9　区域图

图 16.9 为区域图。该图由两个判别函数构成坐标系，其中，函数 1 的不同取值作为横坐标，函数 2 的不同取值作为纵坐标。根据函数值的不同，坐标区域被划分为 3 个部分。如图 16.9 所示，数字“ 1 ”、“ 2 ”和“ 3 ”与边界所围成的区域分别对应于 3 种分类，当函数 1 和函数 2 的取值所决定的点落在其中某个区域中时，该点所对应的个案就属于某类。由表 16.19 可知，北京、上海和广州 3 个城市对应的第 1 函数值为 3.068、0.463 和 2.560，第 2 函数值分别为 1.106、1.236 和 0.636，由图 16.9 可以看出，3 个城市的农民消费水平均属第 1 类。

- 个案统计量

表 16.29　个案统计量

按照案例的统计量

案例数目		最高组					第二最高组			判别式得分	
			(P>d\|G=g)								
	实际组	预测组	p	df	(P>d\|G=g)	到质心的平方 Mahalanobis 距离	组	(P>d\|G=g)	到质心的平方 Mahalanobis 距离	函数 1	函数 2
初始　1	1	1	.365	2	1.000	2.013	2	.000	22.217	3.338	–.2499
2	1	1	.803	2	1.000	.439	2	.000	17.248	3.644	–.1434
3	1	2**	.108	2	.705	4.448	1	.295	6.187	2.570	1.331
4	1	1	.841	2	.905	.346	2	.015	8.671	2.763	–.603
5	1	1	.598	2	1.000	1.030	2	.000	20.059	3.096	–1.516
6	1	1	.450	2	.996	1.395	2	.004	12.600	2.098	–1.915
7	2	2	.730	2	.979	.629	1	.013	9.296	.442	.443
8	2	2	.498	2	.991	1.395	1	.009	10.869	1.503	1.795
9	2	2	.761	2	.986	.545	1	.014	9.108	1.201	1.265
10	2	2	.180	2	.999	3.428	1	.001	16.549	1.608	2.691
11	2	2	.573	2	.963	1.114	1	.023	8.585	.436	.181
12	2	2	.918	2	.994	.171	1	.006	10.543	.873	1.286
13	2	2	.448	2	.859	1.608	3	.140	5.231	–.625	.584
14	2	2	.252	2	.620	2.758	3	.380	3.736	–1.070	.596
15	2	2	.423	2	.985	1.723	3	.015	10.059	–.665	1.908
16	2	2	.838	2	.997	.353	1	.003	11.961	–1.047	1.609
17	3	3	.292	2	.926	2.463	2	.074	7.521	–2.997	–1.054
18	3	3	.828	2	1.000	.378	2	.000	17.276	–1.570	–1.067
19	3	3	.460	2	.991	1.552	2	.009	11.045	–1.482	–1.393
20	3	3	.548	2	.969	1.203	2	03P	8.101	–2.350	–.842
21	3	3	.776	2	.988	.508	2	oP2	9.329	–1.986	.046
22	3	3	.659	2	.963	.835	2	.031	7.342	–4.472	.076
23	3	3	.092	2	1.000	4.777	2	000	30.196	–3.594	–.525
24	3	3	.118	2	.998	4.267	2	.002	`16.467	–3.267	1.152

续表

案例数目		最高组					第二最高组			判别式得分	
			(P>d\|G=g)								
	实际组	预测组	p	df	(P>d\|G=g)	到质心的平方 Mahalanobis 距离	组	(P>d\|G=g)	到质心的平方 Mahalanobis 距离	函数 1	函数 2
25	3	3	.261	2	1.000	2.608	2	.000	25.126	5.598	−2.113
26	未分组的	1	.000	2	1.000	19.350	2	.000	61.659	11.727	−4.075
27	未分组的	1	.000	2	1.000	79.861	2	.000	147.691	3.741	−3.342
28	未分组的	1	.797	2	1.000	.453	2	.000	16.101		−1.079
交叉验证[a] 1	1	1	.208	4	1.000	5.886	2	.000	26.151		
2	1	1	.007	4	.999	14.227	2	.001	28.789		
3	1	2**	.063	4	1.000	8.934	1	.000	30.317		
4	1	1	.123	4	.921	7.254	2	.079	12.160		
5	1	1	.342	4	1.000	4.507	2	.000	23.149		
6	1	1	.613	4	.992	2.678	2	.008	12.259		
7	2	2	.698	4	.986	2.203	1	.021	9.905		
8	2	2	.600	4	.985	2.752	1	.015	11.062		
9	2	2	.336	4	.968	4.556	1	.032	11.370		
10	2	2	.193	4	.997	6.082	1	.003	17.673		
11	2	2	.769	4	.947	1.822	1	.033	8.549		
12	2	2	.921	4	.992	.923	1	.008	10.622		
13	2	3**	.064	4	.540	8.887	2	.460	9.208		
14	2	3**	.352	4	.521	4.424	2	.479	4.593		
15	2	2	.511	4	.971	3.288	3	.029	10.296		
16	2	2	.000	4	.887	32.035	1	.112	36.171		
17	3	3	.406	4	.857	4.002	2	.143	7.584		
18	3	3	.808	4	1.000	1.605	2	.000	17.911		
19	3	3	40 ‘	4	.982	3.457	2	.018	11.491		
20	3	3	739	4	.956	1.980	2	.044	8.125		
21	3	3	.901	4	.984	1.054	2	.016	9.276		
22	3	3	.768	4	.950	1.826	2	.050	7.696		
23	3	3	.096	4	1.000	7.878	2	.000	37.216		
24	3	3	.130	4	.993	7.123	2	.007	17.151		
25	3	3	.282	4	1.000	5.050	2	.000	28.839		

对初始数据来说，平方 Mahalanobis 距离基于典则函数。

对交叉验证数据来说，平方 Mahalanobis 距离基于观察值。

**．错误分类的案例。

a．仅对分析的案例进行交叉验证。在交叉验证中，每个案例都是从按照从该案例以外的所有其他案例派生的函数来分类的。

表 16.29 为个案统计量表。表中包括个案数、实际分组、最高组和第二最高组预测分组信息、判别式得分等。表中带用“ ** ”符号的个案表示该个案为错判的个案。

- 散点图

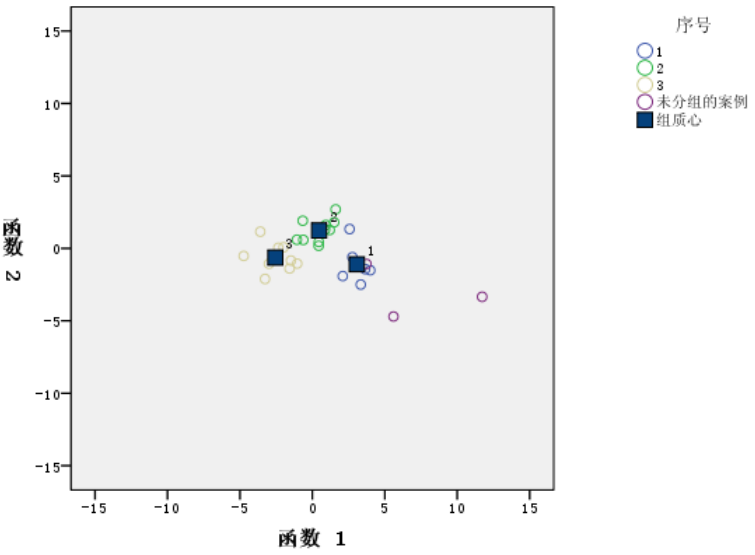


图 16.10　散点图

图 16.10 为所有组散点图。图的横坐标为函数 1，纵坐标为函数 2。图中 3 个已知类别和未分组的个案对应的散点分别用 4 种不同的颜色表示，较大的方框表示 3 个已知类别的质心。通过图 16.10 可以看出，未分组个案距离第一类别的质心最近，所以应该归到第一类中。

- 分类结果

表 16.30　分类结果

序　号			预测组成员			合　计
			1	2	3	
初始	计数	1	5	1	0	6
		2	0	10	0	10
		3	0	0	9	9
		未分组的案例	3	0	0	3
	%	1	83.3	16.7	.0	100.0
		2	.0	100.0	.0	100.0
		3	.0	.0	100.0	100.0
		未分组的案例	100.0	.0	.0	100.0
交叉验证[a]	计数	1	5	1	0	6
		2	0	8	2	10

续表

序　号			预测组成员			合　计
			1	2	3	
		3	0	0	9	9
	%	1	83.3	16.7	.0	100.0
		2	.0	80.0	20.0	100.0
		3	.0	.0	100.0	100.0

a．仅对分析中的案例进行交叉验证。在交叉验证中，每个案例都是按照从该案例以外的所有其他案例派生的函数来分类的。

b．已对初始分组案例中的 96.0%进行了正确分类。

c．已对交叉验证分组案例中的 88.0%进行了正确分类。

表 16.30 为分类结果，从该表中可以判别每个个案是否正确分类。例如，对应于初始计数，6 个个案中有 5 个被正确地判为第 1 类，占 83.3%，有 1 个被错判为第 2 类，占 16.7%。其余各类均判断正确。

16.2　判定新生婴儿的性别比例是否正常

16.2.1　实例内容说明

某地某一时期内出生 35 名婴儿，其中女性 19 名（设 Sex=0），男性 16 名（设 Sex=1）。问这个地方出生婴儿的性别比例与通常的男女性比例（总体概率约为 0.5）是否相同。数据如表 16.31 所示。

表 16.31　35 名新生婴儿性别

婴儿序号	性　别	婴儿序号	性　别
1	1	19	0
2	0	20	0
3	1	21	0
4	1	22	0
5	1	23	1
6	1	24	1
7	0	25	1
8	0	26	1
9	0	27	0
10	0	28	0
11	1	29	0
12	1	30	0

续表

婴儿序号	性　别	婴儿序号	性　别
13	1	31	1
14	1	32	0
15	1	33	0
16	1	34	0
17	0	35	0
18	0		

16.2.2　实现方法分析

非参数检验是指总体分布未知或知之甚少时，利用已知样本数据对总体的分布形态等做出推断的分析方法，由于在推断过程中一般不涉及总体参数而得名。非参数检验主要用于那些总体分布不能用有限个实参数来刻画，或者不考虑被研究的对象为何种分布以及是否已知的情况。

非参数检验的方法非常丰富。根据所要推断的问题或者所利用的统计量，可以分为卡方检验、二项式检验、游程检验、单样本 K-S 检验、两个独立样本检验、多个独立样本检验、两个相关样本检验和多个相关样本检验等。

（1）总体分布的卡方检验适用于配合度检验，是根据样本数据的实际频数推断总体分布与期望分布或理论分布是否有显著差异来检验的。

（2）二项式检验就是根据收集到的样本数据，推断总体分布是否服从某个指定的二项式分布。

（3）游程检验是对某变量的取值是否随机出现进行检验。

（4）单样本 K-S 检验是以两位前苏联数学家 Kolmogorov 和 Smirnov 命名的，也是一种拟合优度的非参数检验方法。单样本 K-S 检验是利用样本数据推断总体是否服从某一理论分布的方法，适用于探索连续型随机变量的分布形态。

（5）两独立样本的非参数检验是在对总体分布不很了解的情况下，通过分析样本数据，推断样本来自的两个独立总体分布是否存在显著差异。一般用来对两个独立样本的均数、中位数、离散趋势、偏度等进行差异比较检验。

（6）多独立样本非参数检验分析样本数据是推断样本来自的多个独立总体分布是否存在显著差异。SPSS 多独立样本非参数检验一般推断多个独立总体的均值或中位数是否存在显著差异。

（7）两相关样本非参数检验是在对总体分布不很清楚的情况下，对样本来自的两相关配对总体分别进行检验。

（8）多相关样本非参数检验是对多个匹配样本的总体分布是否存在显著性差异进行统计分析。SPSS 中有以下 3 种多相关样本非参数检验方法：多相关样本的 Friendman 检验、多相关样本的 Kendall 协同系数检验和多相关样本的 Cochran Q 检验。

本例用 SPSS 分析模块中，非参数检验中的二项式检验来分析出生婴儿的比例是否符合正常比例。

16.2.3 具体操作步骤

在 SPSS 中进行二项式检验，可以利用“ 分析→非参数检验→二项式 ”命令实现，菜单内容如图 16.11 所示。

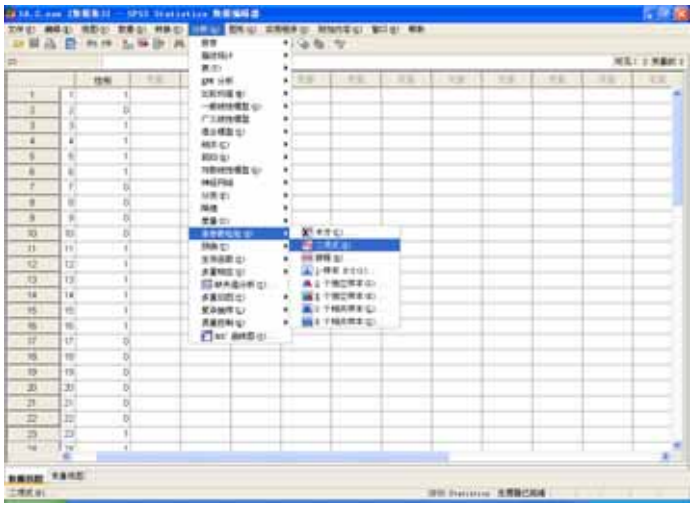


图 16.11 SPSS 二项式检验模块

二项式检验的 SPSS 操作如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“ 分析→非参数检验→二项式 ”，打开“ 二项式检验 ”对话框，如图 16.12 所示。



图 16.12 “ 二项式检验 ”对话框

在对话框的左侧变量列表框中选择“ 性别 ”变量，单击按钮，将其选入到右侧的“ 检验变量 ” 列表框。如图 16.13 所示。

- 二项式检验：精确检验对话框

单击图 16.12 中 “ 精确 ” 按钮，打开 “ 精确检验 ” 对话框，如图 16.14 所示。



图 16.13　变量选入后对话框



图 16.14 “ 精确检验 ” 对话框

“ 精确检验 ” 对话框用于计算显著性水平 Sig.值的几种方法，具体如表 16.32 所示。

表 16.32　精确检验选项及其含义

选　项	含　义
仅渐进法	默认计算显著性水平的方法。计算显著性水平是否是基于检验统计量的渐近分布假设。如果显著性水平小于 0.05，检验结果被认为存在显著性差异。根据大数定理理论，渐近方法的显著性水平要求数据量足够大，如果数据量比较小，或者频数过低，检验结果可能会失效
Monte Carlo	蒙特卡罗估计法，精确显著性水平的无偏估计。蒙特卡罗方法是利用给定样本集通过模拟方法重复取样计算显著性水平，该方法不要求假设。蒙特卡罗方法是一个很有用的方法。对于处理不满足渐近假设的巨量数据，同时由于巨量数据而无法得到精确的显著性水平时，可以选择该方法在 “ 信赖水平 ” 文本框中输入计算显著性水平的置信度，在 “ 样本数 ” 文本框中输入取样数量
精确	精确计算显著性水平的方法。该方法可以得到精确的显著性水平，不需要仅渐近法的假设，不足之处是计算量和所需内存太大。选择该项后，可以选择 “ 每个检验的时间限制为 ” 复选项，即设置限制计算时间，默认时间限制是 5 分钟，超过该时间，系统自动停止计算，给出计算信息

本例选择 “ 仅渐进法 ”。

- 二项式检验：选项对话框

单击图 16.12 中 “ 选项 ” 按钮，打开 “ 二项式检验：选项 ” 对话框。如图 16.15 所示。

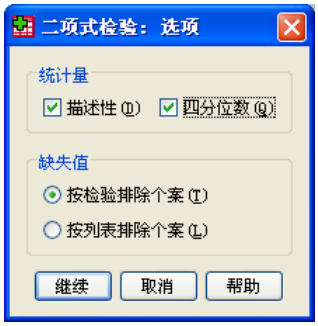


图 16.15　二项式检验：选项对话框

在该对话框的“统计量”选项组中包含选择统计量的两个复选项。其具体含义如表 16.33 所示。

表 16.33　统计选项组选项及含义

选　项	含　义
描述性	描述性统计量，显示均值、标准差、最大值、最小值和非缺失个案数
四分位数	显示 4 分位数，即 25%、50%和 75%的百分位数

在该对话框的“缺失值”选项组中有对缺失值的处理方式，具体如表 16.34 所示。

表 16.34　缺失值选项组选项及含义

选　项	含　义
按检验排除个案	默认选项，在用于检验的个案中，将其中含有缺失值的个案剔除
按列表排除个案	选择该项，表示将含有缺失值的全部个案剔除

03 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 0.
NPAR TESTS
  /BINOMIAL (0.50)=性别
  /STATISTICS DESCRIPTIVES QUARTILES
  /MISSING ANALYSIS.
```

04 结果输出。

“定义二分类法”和“检验比例”选择默认值，设置完毕后，单击“确定”按钮，执行二项式检验。

- 描述性统计量

表 16.35　描述性统计量

描述性统计量

	N	均　值	标准差	极小值	极大值	百分位		
						第 25 个	第 50 个（中值）	第 75 个
性别	35	.46	.505	0	1	.00	.00	1.00

表 16.35 显示的是检验数据的描述性统计量。给出了样本的个案数、均值、标准差、极值（包括极大值和极小值）和百分位等信息。

- 检验结果

表 16.36　二项式检验结果

二项式检验

		类　别	N	观察比例	检验比例	渐近显著性（双侧）
性别	组 P	1	16	.46	.50	.736[a]
	组 2	0	19	.54		
	总数		35	1.00		

a．基于 Z 近似值。

表 16.36 为二项式检验结果。从表中可以看出，第一组的频数为 16，第二组的频数为 19，检验的二项分布的概率值为 0.50，检验统计量为近似正态分布（大样本情况下），计算得到的概率 ρ 值为 0.736，大于 0.05，所以接受零假设，可以认为新生婴儿的性别比例正常。

第17章 SPSS行业应用实例——市场研究/企业数据分析

市场研究/企业数据分析是一种通过将消费者、客户和公众与营销者连接起来的职能，它用于识别和确定市场机会和问题，对营销活动进行策划、改进和评估，对营销绩效进行监测，加深人们对营销过程的理解。市场研究/企业数据分析规定了解决这些问题所需要的，设计收集方法，对收集过程进行管理和实施，对结果进行分析，最后对研究结果及其含义进行沟通。市场研究/企业数据分析有三种功能描述：收集并陈述实事、诊断，或对活动进行解释、预测，分析市场随时间的变化。

17.1 某商品销售量与超市规模和摆放位置的方差分析

17.1.1 实例内容说明

为了研究超市中某种商品的销量与摆放位置和超市规模的关系，按照超市规模（大型、中型、小型三种水平）和摆放位置（A、B、C、D四点）各随机选取了两个点，记录其同一周期内同种商品的销售量，试对其做单变量多因素方差分析。原始数据如表17.1所示。

表17.1 同一周期内同种货物销售量

超市规模	摆放位置							
	A		B		C		D	
大型	70	78	75	82	82	89	71	75
中型	57	65	69	78	73	80	60	57
小型	45	50	56	63	65	71	48	53

17.1.2　实现方法分析

因受不同因素的影响，研究所得的数据会不同。造成结果差异的原因可分成两类：一类是受不可控的随机因素的影响，这是人为很难控制的一类影响因素，称为随机变量；另一类是研究中人为施加的可控因素对结果的影响，称为控制变量。方差分析就是实现上述功能的分析方法。方差分析是 R.A.Fister 发明的，用于两个及两个以上样本均数差别的显著性检验。方差分析方法在不同领域的各个分析研究中都得到了广泛的应用。从方差入手的研究方法有助于找到事物的内在规律性。

方差分析的基本思想是：通过分析研究不同变量的变异对总变异的贡献大小，确定控制变量对研究结果影响力的大小。通过方差分析不同水平的控制变量是否对结果产生了显著影响。如果控制变量的不同水平对结果产生了显著影响，那么它和随机变量共同作用，必然使结果有显著的变化；如果控制变量的不同水平对结果没有显著的影响，那么结果的变化主要由随机变量起作用，和控制变量关系不大。

根据控制变量的个数，可以将方差分析分成单因素方差分析、多因素方差分析和协方差分析。

（1）单因素方差分析测试某一个控制变量的不同水平是否给观察变量造成了显著差异和变动。单因素方差分析实质上采用了统计推断的方法，由于方差分析有一个比较严格的前提条件，即不同水平下，各总体均值服从方差相同的正态分布，因此方差分析问题就转换成了研究不同水平下各个总体的均值是否有显著差异的问题。

（2）多因素方差分析中的控制变量在两个或两个以上，它的研究目的是要分析多个控制变量的作用、多个控制变量的交互作用以及其他随机变量是否对结果产生了显著影响。多因素方差分析不仅需要分析多个控制变量独立作用对观察变量的影响，还要分析多个控制变量交互作用对观察变量的影响，及其他随机变量对结果的影响。因此，它需要将观察变量总的离差平方和分解为 3 个部分：多个控制变量单独作用引起的平方和、多个控制变量交互作用引起的离差平方和、其他随机因素引起的离差平方和。

（3）协方差分析是将那些很难控制的因素作为协变量，在排除协变量影响的条件下，分析控制变量对观察变量的影响，从而更加准确地对控制因素进行评价。利用协方差分析就可以完成这样的功能。协方差将那些很难控制的随机变量作为协变量，在分析中将其排除，然后再分析控制变量对观察变量的影响，从而实现对控制变量效果的准确评价。协方差分析要求协变量应是连续数值型，多个协变量间相互独立，且与控制变量之间没有交互影响。

单因素方差分析所解决的是一个因素下的多个不同水平之间的相关问题；多因素方差分析的控制变量在两个或两个以上，其主要用于分析多个控制变量的作用、多个控制变量的交互作用以及其他随机变量是否对结果产生了显著影响；协方差分析将那些很难控制的因素作为协变量，在排除协变量影响的条件下，分析控制变量对观察变量的影响，从而更

准确地对控制因素进行评价。

在本例中所关心的变量是销售量，可能对其有影响的因素有两个：超市规模和货物摆放位置，因此本例问题可以使用 SPSS 方差分析模块中的单变量多因素方差分析来解决。

17.1.3 具体操作步骤

根据前面分析，本例问题需要用方差分析进行解决，因此，应该选择一般线性模型中的单变量多因素方差分析。

在 SPSS 中进行单变量方差分析，可以利用“ 分析→一般线性模型→单变量 ”命令实现，菜单内容如图 17.1 所示。

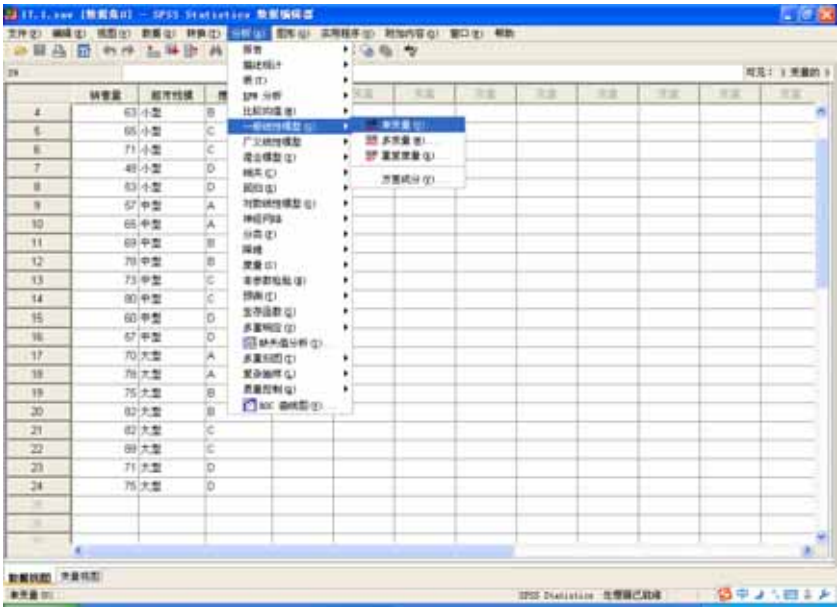


图 17.1　SPSS 单变量多因素方差分析模块

单变量多因素方差分析的 SPSS 操作如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“ 分析→一般线性模型→单变量 ”命令，打开“ 单变量方差分析 ”对话框。如图 17.2 所示。

在对话框的左侧变量列表框中选中变量“ 销售量 ”，单击 按钮，将其移动到右侧的“ 因变量 ”列表框。用同样方法将变量“ 超市规模 ”和“ 摆放位置 ”选入“ 固定因子 ”列表框。结果如图 17.3 所示。



图 17.2 “ 单变量 ” 方差分析对话框

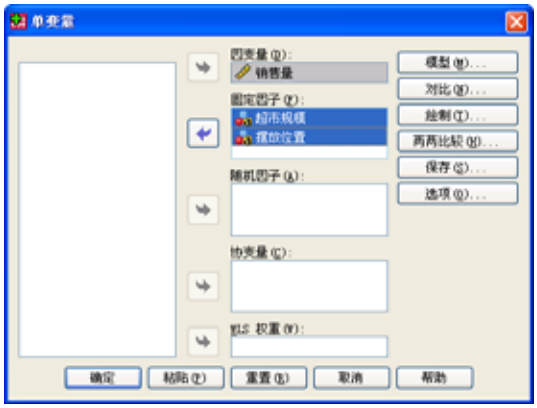


图 17.3　选入变量后对话框

- 单变量：模型对话框

单击图 17.2 中 “ 模型 ” 按钮，打开 “ 单变量：模型 ” 对话框。如图 17.4 所示。

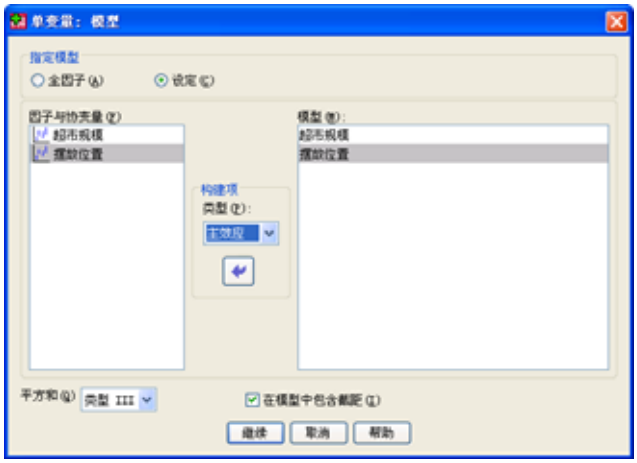


图 17.4 “ 单变量：模型 ” 对话框

该对话框用于设定计算的模型等参数。

“ 指定模型 ” 选项组指定模型的类型，包括表 17.2 中的两个选项。

表 17.2　指定模型各选项意义

选　　项	意　　义
全因子	默认选项。全因子模型包括所有因素主效应、所有协变量主效应及所有因素与因素的交互效应，不包括协变量的交互效应
设定	只指定一个交互效应的子集或因素与协变量的交互效应。用户必须指定模型的所有项目

如果在 “ 指定模型 ” 选项组选择了 “ 设定 ” 选项，在 “ 因子与协方差 ” 变量列表框中

自动列出可以作为因素变量的变量名，名称后用括号内的字母 F 标识；也列出可以作为协变量的变量名，名称后用括号内的字母 C 标识。

单击“ 构建项 ”中的下拉箭头，在弹出的下拉菜单中选择要构建条件的选项，如表 17.3 所示。

表 17.3　构建项各选项及其含义

选　　项	含　　义
交互	建立所有被选变量最高水平的交互效应项目，默认选项
主效应	建立每个被选变量主效应
所有 2 阶	建立所有 2 阶交互效应
所有 3 阶	建立所有 3 阶交互效应
所有 4 阶	建立所有 4 阶交互效应
所有 5 阶	建立所有 5 阶交互效应

该对话框“ 平方和 ”下拉列表用于选择平方和的分解方法，单击下拉按钮，出现下拉菜单，各选项及含义如表 17.4 所示。

表 17.4　构建项各选项及其含义

选　　项	含　　义
类型 I	Ⅰ型平方和，适用于平衡方差分析模型、多项式回归模型、完全嵌套设计
类型 II	Ⅱ型平方和，适用于平衡方差分析模型、只有一个主因子的模型、回归模型、完全嵌套设计
类型III	Ⅲ型平方和，适用于Ⅰ型、Ⅱ型平方和适用的所有模型、无缺失值的所有平衡与不平衡模型。此项较常用，为默认选项
类型IV	Ⅳ型平方和，适用于Ⅰ型、Ⅱ型平方和适用的所有模型、无缺失值的所有平衡与不平衡模型

“ 在模型中包含截距 ”选项表示的含义是：通常截距包括在模型汇总，如果能假设数据通过原点，可以不包括截距，即不选择此项。

本例中“ 指定模型 ”选项组选择“ 设定 ”；将“ 因子与协变量 ”列表框中的变量选入到“ 模型 ”列表框中；“ 平方和 ”选项组选择“ 类型III ”；“ 构建项 ”类型选择“ 主效应 ”。



图 17.5　“ 单变量：对比 ”对话框

● 单变量：对比对话框

单击图 17.2 中“ 对比 ”按钮，打开“ 单变量：对比 ”对话框。如图 17.5 所示。

在“ 因子 ”列表框中显示出所有在主对话框中选中的因素变量，因素变量名后的括号是当前的对比方法。可以在“ 更改对比 ”选项区域中改变对比方法，方法是“ 因子 ”列表框中选择需要改变对比方法的因子，然后单击“ 对比 ”下拉菜单按钮，在弹出的下拉菜单中选择如下对比方法，如表 17.5 所示。

表 17.5　对比方法各选项及其含义

方　法	含　义
无	不进行均值比较
偏差	差别对比，比较每一水平的平均效应和总平均效应选择
简单	简单对比，每个水平与指定水平的平均效应比较，可选择第一组和最后一组作为参考水平
差值	差分对比，每个水平与前面各水平的平均效应的比较
Helmert	Helmert 对比，每个水平与随后的各水平的平均效应的比较
变量	重复对比，每个水平与随后的一个水平的比较
多项式	多项式对比，比较线行效应、二次效应、三次效应等，常用于预测多项式趋势

改变对比的设置完成后，单击“ 更改 ”按钮，改变了对照方法显示在“ 因子 ”列表框中。设置完毕后，单击“ 继续 ”按钮，返回到主对话框。

- 单变量：轮廓图对话框

单击图 17.2 中“ 绘制 ”按钮，打开“ 单变量：轮廓图 ”对话框。如图 17.6 所示。



图 17.6 “ 单变量：轮廓图 ”对话框

在该对话框中可以绘制一个或多个因素变量为参考的因变量均值分布图。在只有一个因素水平时，为因变量估计边缘均值的轮廓图；在两个以上因素水平时，绘制分离线。

“ 因子 ”列表框显示因素的变量名，将该列表框中的因素变量移动到右侧的不同列表框中，可选列表框主要有：水平轴、单图和多图。

本例中，将“ 因子 ”列表框的两个变量分别移动到“ 水平轴 ”窗口后，单击“ 添加 ”按钮，将其选入到“ 图 ”列表框。

- 单变量：两两比较对话框

单击图 17.2 中所示“ 两两比较 ”按钮，打开“ 单变量：观测均值的两两比较 ”对话框。如图 17.7 所示。



图 17.7 “ 单变量：观测均值 ” 的两两比较对话框

“ 假定方差齐性 ”选项组给出当方差相等时的确定多重比较方法，包括 14 种可选方法，具体如表 17.6 所示。

表 17.6 假定方差齐性各选项及其含义

方 法 名	含 义
LSD	Least-significant difference，即最小显著性差异法。用 T 检验完成个组间的配对比较，检验的敏感性高，各个水平间的均值存在微小的差异也有可能被检验出来，但此方法对第一类弃真错误的概率不进行控制和调整
Bonferron i	LSDMOD 方法，即修正最小显著差异法。用 T 检验完成各组间的配对比较；同时，通过设置每个检验的误差率来控制第一类错误的概率
Sidak	Sidak 法，根据 *t* 统计量进行配对多重比较，调整多重比较的显著性水平，其界限比 Bonferroni 方法小
Scheffe	Scheffe 差别检验法。α 可指定 0~1 之间任何显著性水平，默认值为 0.05。
R-E-G-W-F	Ryan-Einot-Gabriel-Welsch F 方法，根据 *F* 统计量检验的 Ryan-Einot-Welsch 多重比较
R-E-G-W-Q	Ryan-Einot-Gabriel-Welsch Q 方法，根据 Student 极差统计量的 Ryan-Einot-Welsch 多重比较
S-N-K	Student-Newman-Keuls 方法，根据 Student 极差统计量在均值间进行配对比较；如果各组样本含量相等或者选择了 “ Harmonic average of all groups ” (所有个组样本含量的调和平均值)，即用所有个组样本含量的调和平均值进行样本量的估计时，还用逐步过程进行齐次子集(差异较小的子集)的均值配对比较；在该比较过程中各组均值按从大到小的顺序排列，最先比较最末端的差异
Tukey	Tukey’s honestly significant difference 方法，即 Tukey 显著差异法，用 Student-Range 统计量进行所有组间均值的配对比较，用所有配对比较误差率作为实验误差率
Tukey’s-b	Tukey-b 方法，用 Student-Range 分布进行组间均值的配对比较，其精确值为两种检验相应值的平均值

续表

方 法 名	含　义
Duncan	Duncan 方法，指定一系列的 Range 值逐步进行计算比较得到结论
Hochberg's GT2	Hochberg's GT2 方法，用正态最大系数进行多重比较
Gabriel	Gabriel 方法，用正态标准系数进行配对比较，单元较大时这种方法较自由
Waller-Duncan	Waller-Duncan 方法，用 t 统计量进行多重比较检验，使用贝叶斯过程的多重比较检验
Dunnett	Dunnett 方法，指定此方法进行配对比较，方法是选择最后一组为对比其他个组和它比较，选定此方法后激活“ Control Catetory ”栏，在该栏展开下拉菜单选择对比组，其中可选“ Last ”分类和“ First ”分类；并在“ Test ”选项组中设定单双侧检验的选择，其中可选“ 2-Sided、<Control、>Control ”

“ 未假定方差齐性 ”选项组中给出当方差不齐时的多重比较方法选项，包括如表 17.7 中的 4 项。

表 17.7　未假定方差齐性各选项及其含义

方 法 名	含　义
Tamhane's T2	用 T 检验进行配对比较试验
Dunnett's T3	用基于 Student 最大模数的比较配对试验
Games Howell	方差不齐时的配对比较试验
Dunnett's C	用基于 Student 极差的配对比较试验

本例中将“ 因子 ”列表框中的变量选入到“ 两两比较检验 ”列表框；“ 假定方差齐性 ”选项组选择“ S-N-K ”。

- 单变量：保存对话框

单击图 17.2 中所示“ 保存 ”按钮，打开“ 单变量：保存 ”对话框。如图 17.8 所示。

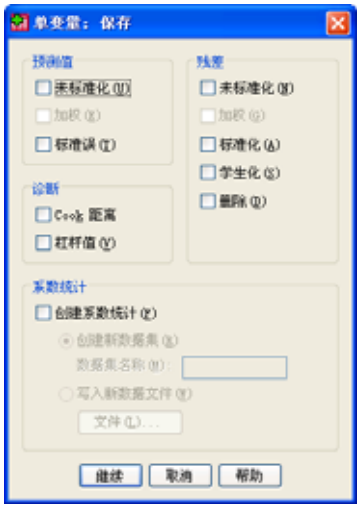


图 17.8　“ 单变量：保存 ”对话框

该对话框给出预测值残差和检测值作为新的变量的保存选项。

“ 预测值 ”选项组各个选项的含义如表 17.8 所示。

表 17.8　预测值各选项及其含义

选　项	含　义
未标准化	非标准化预测值
加权	加权非标准化预测值。如果在主对话框中选择 WLS（加权最小二乘）变量，保存因变量的加权非标准化预测值
标准误差	预测值均值的标准误差

“ 诊断 ”选项组给出诊断结果保存选项，如表 17.9 所示。

表 17.9　诊断各选项及其含义

选　项	含　义
Cook’s 距离	Cook 距离
杠杆值	非中心化杠杆值

“ 残差 ”选项组给出诊断结果保存选项，如表 17.10 所示。

表 17.10　残差各选项及其含义

选　项	含　义
未标准化	非标准化残差值。残差为观测值与预测值之差
加权	权重非标准化残差。如果在主对话框中选择 WLS 变量，则保存权重非标准化残差
标准化	标准化残差，即 pearson 残差
学生化	学生化残差
删除	删除残差自变量值与调整预测值之差

本例选择默认参数。

- 单变量：选项对话框

单击图 17.2 中所示“ 选项 ”按钮，打开“ 单变量：选项 ”对话框。如图 17.9 所示。



图 17.9 “ 单变量：选项 ”对话框

“ 估计边际均值 ”选项组用于选择显示结果的因素变量，方法是在左侧的“ 因子和因子交互 ”列表框中选择因素和交互项，单击右向箭头按钮，将其移动到“ 显示均值 ”列表框。

如果在“ 显示均值 ”栏中主效应变量被选中，则可以选择“ 比较主效应 ”选项，对主效因变量进行估计边际均值，并可在“ 置信区间调节 ”下拉菜单中选择“ LSD ”、“ Bonferroni、Sidar ”方法。

“ 输出 ”选项组指定要求输出的统计量，其各选项及含义如表 17.11 所示。

表 17.11　显示各选项及其含义

选项	含　义
描述统计	输出描述统计量。显示所有单元格中因变量的观测均值、标准差和例数
功效估计	效应大小的估计。计算每个效应的估计及参数估计、标准误差、置信区间
检验效能	给出计算功效的显著性水平 Alpha 值，该值在 0.01~0.99 之间。显示观测功效系统，默认显著性水平是 0.05
参数估计	给出各因变量与自变量的回归
对比系数矩阵	输出变量之间的对比系数矩阵
方差齐性检验	输出方差齐次检验结果
分布-水平图	可绘制观测值与残差的散点图
残差图	绘制残差图
缺乏拟合优度检验	拟合度不足的检验，检查独立变量和非独立变量间的关系是否被充分描述
一般估计函数	可以根据一般估计函数自定义假设检验，对比系数矩阵的行与一般函数的线性组合

在“ 显著性水平 ”文本框中输入显著性水平，用于检验后多重比较的置信水平和计算检验的观察效能。

本例“ 输出 ”选项组选择“ 方差齐性检验 ”。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
UNIANOVA 销售量 BY 超市规模 摆放位置
  /CONTRAST(超市规模)=Simple
  /CONTRAST(摆放位置)=Simple
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /POSTHOC=超市规模 摆放位置(SNK)
  /PLOT=PROFILE(超市规模 摆放位置)
  /PRINT=HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=超市规模 摆放位置.
```

04 结果输出。

设置完毕后，单击“ 确定 ”按钮，执行单变量方差分析。

- 主体间因子

表 17.12　主体间因子

主体间因子

		N
超市规模	大型	8
	小型	8
	中型	8
摆放位置	A	6
	B	6
	C	6
	D	6

表 17.12 给出主体的因子数。变量“超市规模”有 3 个水平，即大型、中型和小型，每个水平有 8 个个案；变量“摆放位置”有 4 个水平，即 A、B、C 和 D，每个水平有 6 个个案。

- 误差方差等同性的 Levene 检验

表 17.13　误差方差等同性的 Levene 检验

误差方差等同性的 Levene 检验[a]

因变量：销售量

F	df1	df2	Siq
	11	12	

检验零假设，即要所有组中因变量的误差方差均相等。

a．设计：截距+超市场规模+摆放位置。

表 17.13 为方差齐性检验的输出，可见无法得出分析结果。这是因为两个因素的各水平较差，一共会形成 12 个单元格，这里检验的就是这 12 个单元格的方差是否齐性。但如果要在考虑交互作用的模型中进行方差齐性检验，每个单元格内至少要有 3 个元素（数据点）才可以，因此这里无法得到分析结果。可见多因素时方差齐性检验往往价值不大。

- 主体间效应的检验

表 17.14　主体间效应的检验

主体间效应的检验

因变量：销售量

| 源 | ||型平方和 | df | 均　方 | F | Siq. |
|---|---|---|---|---|---|
| 校正模型 | 2930.417[a] | 5 | 586.083 | 30.409 | .000 |
| 截距 | 108272.667 | 1 | 108272.667 | 5617.799 | .000 |
| 超市规模 | 1828.083 | 2 | 914.042 | 47.426 | .000 |

续表

| 源 | ||| 型平方和 | df | 均　方 | F | Siq. |
|---|---|---|---|---|---|
| 摆放位置 | 1102.333 | 3 | 367.444 | 19.065 | .000 |
| 误差 | 346.917 | 18 | 19.273 | | |
| 总计 | 111550.000 | 24 | | | |
| 校正的总计 | 3277.333 | 23 | | | |

a．R 方=.894（调整 R 方=.865）

表 17.14 为主体间效应的检验，即去除了交互项的方差分析表。第一行的校正模型是对所用方差分析模型的检验，其原假设对模型中所有影响因素均无作用，即超市规模、摆放位置、两者之间的交互作用均对销售量无影响；第二行是对模型中常数项是否等于 0 进行的检验，显然它在分析中没有实际意义，忽略即可；第三、四行分别是对超市规模、摆放位置的效应进行的检验；第五行是误差分析。

从表中可以看出，同种商品不同规模超市和不同摆放位置的“销售量”的检验统计量 f 的观测值为 30.409，检验的概率 ρ 值为 0.000，小于 0.05，拒绝零假设，可以认为同种商品在不同规模超市和不同摆放位置的情况下，销售量存在显著差异。

- 销售量

表 17.15　销售量

（a）不同超市规模

销售量

Student-Newman-Keuls[a..b]

超市规模	N	子　集		
		1	2	3
小型	8	56．38		
中型	8		67.38	
大型	8			77.75
Sig.		1.000	1.000	1.000

已显示同类子集中的组均值。

基于观测的均值。

误差为均值方（错误）=19.273。

a．使用调和均值样本大小=8.000。

b．Alpha=.05。

（b）不同摆放位置

销售量

Student-Newman-Keuls[a..b]

摆放位置	N	子　集		
		1	2	3
D	6	60.67		
A	6	60.83		

续表

摆放位置	N	子　集		
		1	2	3
B	6		70.50	
C	6			76.67
Siq.		.948	1.000	1.000

已显示同类子集中的组均值。

基于观测的均值。

误差为均值方（错误）=19.273。

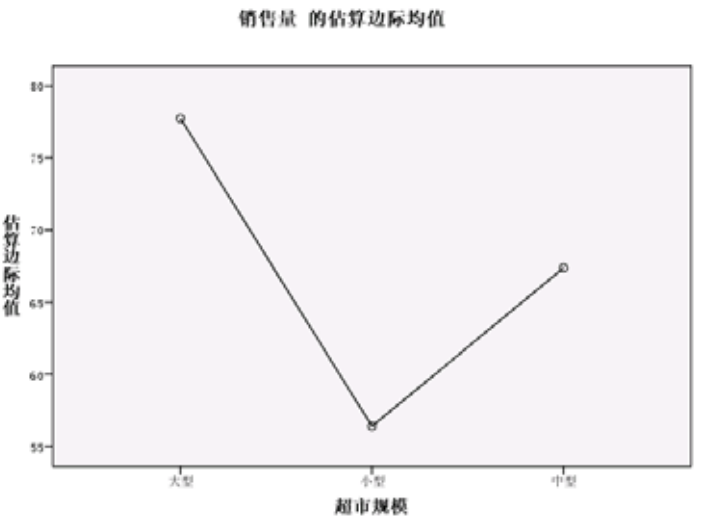
a．使用调和均值样本大小=6.000。

b．Alpha=.05。

表 17.15（a）（b）分别是对超市规模和摆放位置水平间差异的检验，可见超市规模越大，相应的销售量就越大；而 4 种摆放位置也对销售量有影响，C 位置的销售量最大，B 位置其次，A 位置和 D 位置的销售量最小。同时，以上差异不受另一个因素水平取值的影响，即两者间无交互作用。

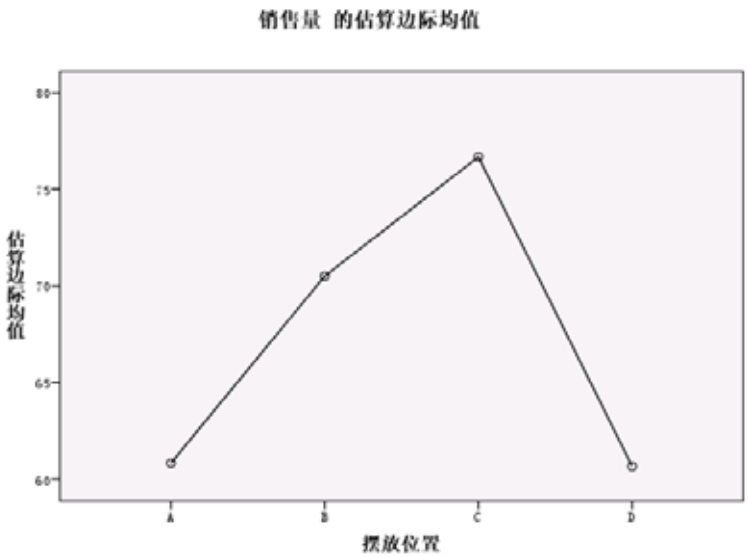
- 轮廓图

图 17.10 为超市规模和摆放位置的轮廓图。轮廓图用于比较各种水平组合下均值的变化规律，发现可能存在的交互作用非常有用。如果两个因素间无交互作用，则第一因素各水平间均值的差异不会随着第二个因素水平取值的变化而变化，表现为绘制的各条曲线基本平行；反之，如果各条曲线明显不平行，之间有剧烈的交叉，则这两个因素的相应水平间可能存在着交互作用。



(a) 超市规模轮廓图

图 17.10　轮廓图

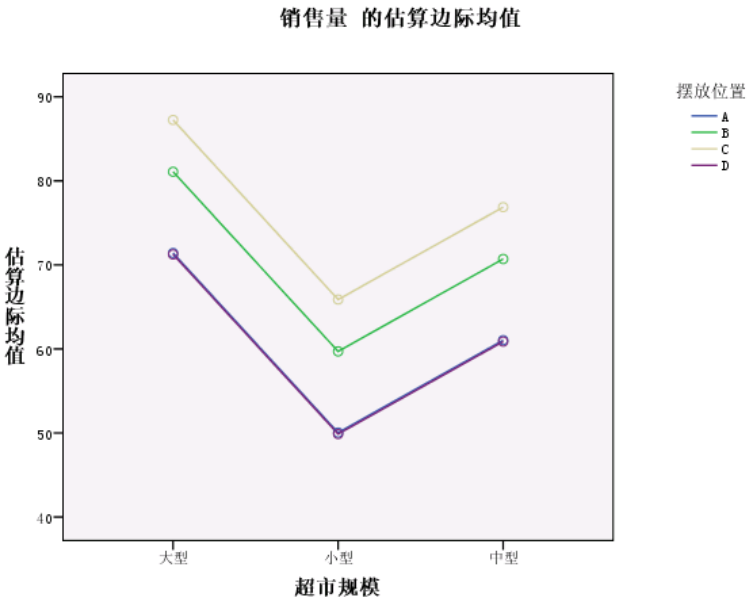


（b）摆放位置轮廓图

图 17.10　轮廓图（续）

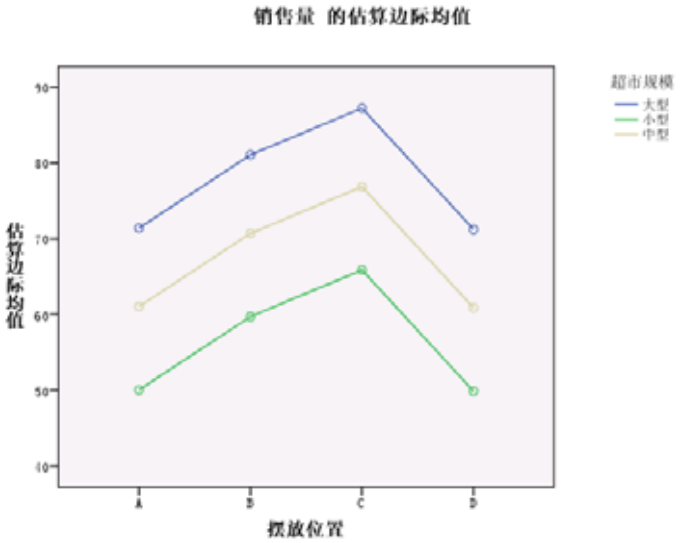
- 加入交互作用的轮廓图

在图 17.6“ 单变量：轮廓图 ”对话框中，将一个变量选入“ 水平轴 ”，另一个变量选入“ 单图 ”，单击“ 添加 ”按钮，将其选入“ 图 ”列表框。得到加入交互作用的轮廓图。



（a）加入交互作用的超市规模轮廓图

图 17.11　加入交互作用的轮廓图



(b) 加入交互作用的摆放位置轮廓图

图 17.11　加入交互作用的轮廓图（续）

图 17.11 为加入交互作用的轮廓图。由图可知，当模型中有交互项时，各边际均值实际上和样本单元格均值相同，且加入交互作用的条曲线大致平行，并未出现明显的交叉，这说明两变量间的交互作用不明显。

- 残差图

在图 17.9 “ 单变量：选项 ” 对话框中，选中 “ 输出 ” 选项组的 “ 残差图 ” 选项，得到执行结果如图 17.12 所示。

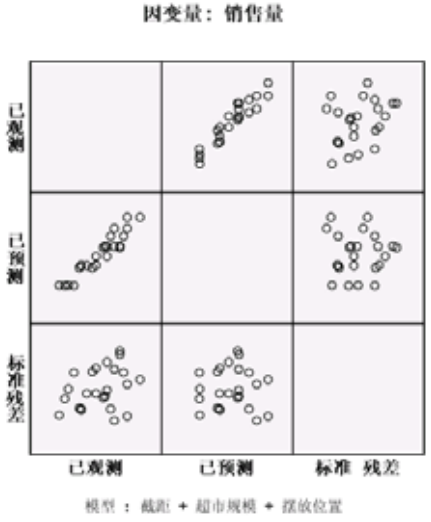


图 17.12　残差图

图 17.12 为残差图。这幅图实际上是一个散点图矩阵，由已观测、已预测和标准残差构成，如果模型拟合效果很好，则已预测值和已观测值应当有明显的相关，呈现出较好的直线趋势，而标准残差则应当完全随机地分布在 0 周围，不随已预测值的上升而出现变化。由图 17.12 中的已预测：标准残差单元格可见，残差的分布的确较好，未发现明显违反模型中正态分布假设的情况，因此该模型的拟合效果是令人满意的。

17.2　汽车配件公司某产品尺寸的参数检验分析

17.2.1　实例内容说明

一个生产高性能汽车配件的公司生产直径为 322mm 的圆盘制动闸。公司的质量控制部门随机抽取不同机器生产的制动闸进行检验。共有 4 台机器，每台机器抽取 16 支产品。表 17.16 给出测量结果。试检验每台机器生产的产品均值和 322 在 90%的置信水平下是否有显著差异。

表 17.16　不同机器生产的圆盘制动闸直径测量数据

机器编号	圆盘制动闸直径
机器 1	322.0003 322.0048 322.0215 321.9907 322.0109 321.9954 322.0059 321.9759 321.9981 321.9957 321.9841 321.9836 322.0037 322.0000 322.0030 322.0023
机器 2	322.0069 322.0306 322.0114 322.0289 322.0087 322.0257 322.0176 322.0070 322.0183 321.9864 322.0178 322.0182 322.0200 322.0122 322.0137 322.0048
机器 3	321.9863 321.9899 322.0016 321.9843 322.0174 321.9834 322.0024 322.0005 322.0039 322.0161 322.0017 322.0085 321.9902 321.9932 321.9912 322.0019
机器 4	321.9939 322.0030 322.0055 322.0028 321.9863 322.0017 321.9978 321.9914 321.9957 321.9989 321.9898 322.0023 321.9861 321.9910 321.9825 321.9982

17.2.2　实现方法分析

在研究实际问题的过程中，通常对研究对象的全部特征或部分特征不是很清楚，如果将研究对象看成一个总体，则可以从总体中抽取样本数据，利用样本数据和已知的部分信息对总体的未知特征进行统计推断，这里的未知特征可以是总体的分布，也可以是总体的某些参数，这就是统计学中的统计推断问题。

统计推断主要有两种方法：

第一种当总体分布已知时，对总体的参数进行推断。例如，当已知总体服从正态分布，需要对正态分布的均值和方差进行推断，在一定的置信度条件下估计参数的取值范围，或者在一定的显著性水平下，对给定参数的取值进行检验。一般情况下，上述总体已知的情况下的统计推断问题被称之为参数检验问题。参数检验不仅可以针对一个总体的检验，也可以针对两个或多个总体比较。

参数检验包括单样本 T 检验、独立样本 T 检验和配对样本 T 检验。单样本 T 检验用于检验正态总体的均值是否与给定的检验值之间存在显著性差异，要求样本数据来自于服从正态分布的单总体；独立样本 T 检验用于检验两个独立样本是否来自于具有相同均值的总体，也就是检验两个独立正态总体的均值是否相等；配对样本 T 检验用于检验两个相关样本是否来自于具有相同均值的总体，即对于两个配对样本，推断两个总体的均值是否存在显著差异。

第二种当总体分布未知时，需要根据样本推断总体的分布类型和参数值的大小，这类总体分布未知的统计推断过程称为非参数检验。

由于该问题设计的是 4 个单总体，而且要进行总体均值检验，同一机器生产的产品直径可近似认为服从正态分布，因此，本例可以利用 SPSS 的单样本 T 检验过程进行分析。

17.2.3 具体操作步骤

在 SPSS 中进行单样本 T 检验，可以利用“ 分析→比较均值→单样本 T 检验 ”命令实现，菜单内容如图 17.13 所示。

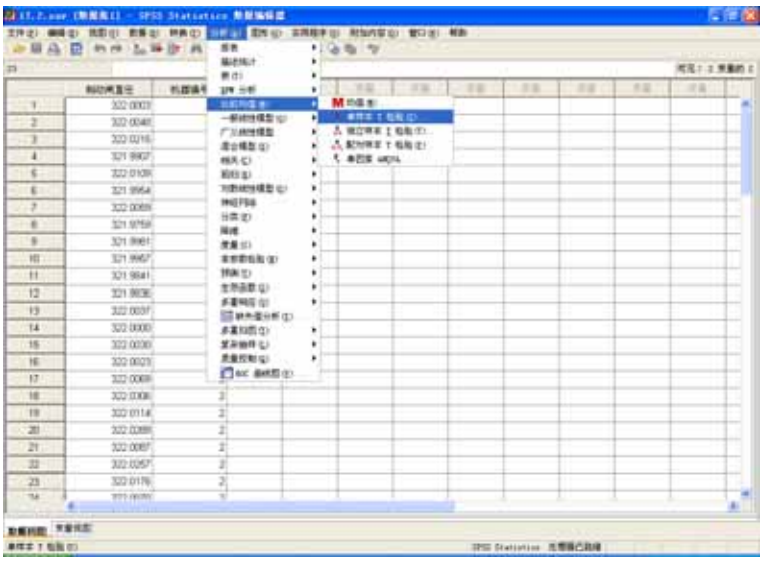


图 17.13 SPSS 单样本 T 检验模块

单样本 T 检验的 SPSS 操作如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 拆分数据文件。

“ 机器编号 ”为定类型变量，用以区别不同机器生产的制动闸。因为对不同机器生产的产品分别进行检验，所以必须先利用变量“ 机器编号 ”对数据文件进行拆分。菜单内容

如图 17.14 所示。

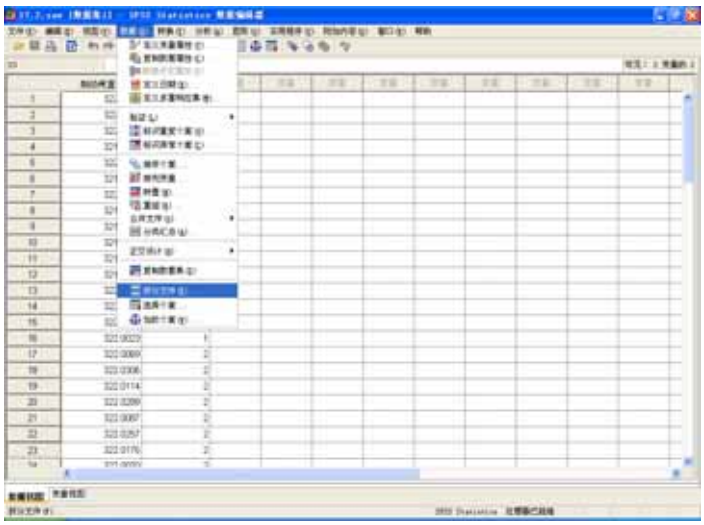


图 17.14　SPSS 拆分文件模块

依次单击菜单“ 数据→拆分文件 ”命令，打开“ 拆分文件 ”对话框，如图 17.15 所示。

在图 17.15 所示对话框中，选择“ 比较组 ”，再将左侧变量列表框中的变量“ 机器编号 ”移动至右侧“ 分组方式 ”列表框中，如图 17.16 所示。



图 17.15　“ 拆分文件 ”对话框



图 17.16　变量选入后的对话框

单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
SORT CASES  BY 机器编号.
SPLIT FILE LAYERED BY 机器编号.
```

单击“ 确定 ”按钮，执行数据拆分。

03 单样本 T 检验选入变量及参数设置。

依次单击菜单“ 分析→比较均值→单样本 T 检验 ”命令，打开“ 单样本 T 检验 ”对话

框，如图 17.17 所示。



图 17.17　单样本 T 检验对话框

在对话框的左侧变量列表框中选择“ 制动闸直径 ”变量，单击按钮，将其选入到右侧的“ 检验变量 ”列表框，如图 17.18 所示。



图 17.18　变量选入后的对话框

将“ 检验值 ”设置为 322。

- 单样本 T 检验：选项对话框

单击图 17.14 中“ 选项 ”按钮，打开“ 单样本 T 检验：选项 ”对话框。如图 17.19 所示。

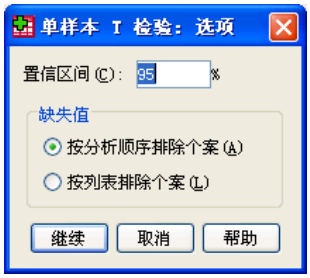


图 17.19 “ 单样本 T 检验：选项 ”对话框

该对话框用于指定置信区间和缺失值的处理方法。

在“信赖区间”文本框中输入置信水平，默认置信水平为 95%，也可以输入 1~99 之间的任意数字。在“缺失值”选项组中指定缺失值的处理方式，有两个单选项，分别为：“排除分析中有缺失值的观测量”和“排除因变量或自变量有缺失值的观测量”。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 1.
T-TEST
  /TESTVAL=322
  /MISSING=ANALYSIS
  /VARIABLES=制动闸直径
  /CRITERIA=CI(.95).
```

05 单样本 T 检验结果输出。

单击图 17.18 对话框中的“确定”按钮，执行单样本 T 检验。

- 单个样本统计量

表 17.17　单个样本统计量

单个样本统计量

机器编号		N	均　值	标 准 差	均值的标准误
1	制动闸直径	16	321.998494	.0111445	.0027861
2	制动闸直径	16	322.014263	.0106985	.0026746
3	制动闸直径	16	321.998281	.0104840	.0026210
4	制动闸直径	16	321.995431	.0069878	.0017469

表 17.17 给出的是单样本 T 检验的描述性统计量，包括每一个样本的均值、标准差和均值的标准误差，均值的标准误差的计算是标准差和样本容量开方的比值。

- 单个样本检验

表 17.18　单个样本检验

机器编号		检验值=322					
						差分的 95%置信区间	
		t	df	Siq.（双侧）	均值差值	下限	上限
1	制动间直径	–.541	15	.597	–.0015063	–.007445	.004432
2	制动闸直径	5.333	15	.000	.0142625	.008562	.019963
3	制动闸直径	–.656	15	.522	–.0017P87	–.007305	.003868
4	制动闸直径	–2.615	15	.019	–.0045688	–.008292	.000845

表 17.18 是单个样本 T 检验的结果列表。表中第二列是 t 统计量的观测值，分别为 -0.541、5.333、-0.656 和-2.615；第三列 df 表示自由度，t 统计量的自由度为样本容量减 1，每一个样本的容量都为 15，故自由度为 15（16-1）；第四列 $Sig.(2-tailed)$ 表示统计量的双尾概率 ρ 值；第五列是样本均值和检验值的差；第六列和第七列分别是在置信度为 90%时总体均值与原检验值之差的置信下限和置信上限，共同构成了该差值的置信区间，例如第

一组样本的置信区间为（-0.007445　0.004432）。

置信水平为 90%时，显著性水平为 0.10，第一组的双尾概率 ρ 值为 0.597，大于 0.10，故零假设成立；观察其他样本的概率 ρ 值可以发现，第二组和第四组的概率 ρ 值小于 0.10，第一组和第三组的概率 ρ 值大于 0.10，故可以认为编号为 2 和 4 的机器生产的产品均值与 322 有显著差异，编号为 1 和 3 的机器生产的产品均值为 322。

第18章 SPSS行业应用实例——证券金融统计

简单来说，证券是多种经济权益凭证的统称，用来证明持券人有权按其券面所载内容取得应有权益的书面证明，按其性质不同可将证券分为证据证券，凭证证券和有价证券，人们通常所说的证券即有价证券。而金融就是资金的融通，金融是货币流通和信用活动以及与之相联系的经济活动的总称，广义的金融泛指一切与信用货币的发行、保管、兑换、结算，融通有关的经济活动，甚至包括金银的买卖，狭义的金融专指信用货币的融通。

在国外的证券金融行业中，统计分析是一个成熟和必需的工具。随着中国证券金融市场的不断发展，越来越多的学者将统计分析方法引入到证券业中。而 SPSS 也是这些证券金融工作者经常使用的工具之一。本章将通过两个典型实例介绍 SPSS 证券金融统计分析的原理、流程与方法。

18.1 股票日历效应分析

18.1.1 实例内容说明

表 18.1 为 2005 年 9 月 5 日至 2006 年 12 月 5 日沪深股市共 606 条交易数据，试考察沪深股市是否具有周效应（数据来源：Wind 咨询）。

表 18.1 沪深股市交易记录（部分）

交易日期	交易所	开盘指数	日最高指数	日最低指数	收盘指数	涨跌数	涨跌幅度	成交量	成交额
05/09/05	1.00	1189.06	1196.87	1184.01	1196.22			267000.66	117.49
05/09/06	1.00	1197.23	1201.82	1172.06	1173.32	-22.90	-1.91	336852.72	153.05
05/09/07	1.00	1170.77	1194.59	1168.58	1194.31	20.99	1.79	286028.52	131.96

续表

交易日期	交易所	开盘指数	日最高指数	日最低指数	收盘指数	涨跌数	涨跌幅度	成交量	成交额
05/09/08	1.00	1196.27	1201.46	1184.79	1194.41	0.10	0.01	305280.56	145.09
05/09/09	1.00	1194.49	1204.46	1184.42	1189.63	-4.79	-0.40	221557.46	101.46
05/09/12	1.00	1189.75	1191.93	1181.43	1188.21	-1.41	-0.12	170966.88	77.41
05/09/13	1.00	1188.25	1207.40	1186.22	1207.16	18.95	1.59	259137.04	117.18
05/09/14	1.00	1209.27	1217.41	1204.55	1217.26	10.10	0.84	322437.60	149.82
05/09/15	1.00	1217.52	1218.94	1210.72	1217.28	0.02	0.00	272174.92	128.97
05/09/16	1.00	1217.42	1217.76	1208.56	1212.95	-4.33	-0.36	223972.24	104.63
05/09/19	1.00	1213.49	1220.97	1210.19	1220.63	7.68	0.63	246568.92	116.36
05/09/20	1.00	1222.26	1223.56	1204.87	1212.62	-8.01	-0.66	323006.92	151.30
05/09/21	1.00	1210.85	1210.85	1186.98	1187.99	-24.63	-2.03	317994.64	142.57
05/09/22	1.00	1185.69	1185.69	1147.19	1159.96	-28.03	-2.36	286022.74	131.85
05/09/23	1.00	1157.30	1164.60	1149.74	1151.98	-7.98	-0.69	187743.94	85.82
05/09/26	1.00	1150.93	1155.60	1147.27	1155.04	3.06	0.27	145381.33	68.63
05/09/27	1.00	1155.05	1157.80	1132.53	1133.24	-21.80	-1.89	180936.28	85.31
05/09/28	1.00	1131.32	1142.49	1129.05	1131.77	-1.46	-0.13	147515.43	68.31
05/09/29	1.00	1134.44	1155.78	1134.44	1155.48	23.71	2.10	185369.92	86.34
05/09/30	1.00	1156.77	1158.70	1150.63	1155.61	0.13	0.01	144934.81	69.60
05/10/10	1.00	1153.67	1153.67	1137.29	1138.95	-16.66	-1.44	117200.65	54.61
05/10/11	1.00	1137.95	1157.30	1133.90	1157.19	18.24	1.60	134265.60	63.99
05/10/12	1.00	1157.48	1165.67	1152.25	1161.85	4.66	0.40	184672.32	93.35
05/10/13	1.00	1162.02	1165.35	1152.09	1152.61	-9.25	-0.80	169052.40	86.00
05/10/14	1.00	1152.37	1154.32	1135.51	1139.55	-13.06	-1.13	151459.99	77.41
…	…	…	…	…	…	…	….	…	…

18.1.2 实现方法分析

在股票市场上，收益、风险等指标一般都有随着日历变化的特征，称之为日历效应。日历效应可以分为周效应——指标在周内各日表现出不同的特征；月效应——指标在不同月份具有不同的特征。国内外许多学者对价格变动的日历效应进行了大量研究，发现收益率和交易量都存在显著的日历效应。

本例中对股票交易数据进行日历效应分析主要采用 SPSS 中的转换模块中的计算变量、数据中的拆分文件、图形中的条形图绘制工具、因素方差分析和相关分析等。

18.1.3 具体操作步骤

18.1.3.1 计算变量

由于本例要考察的问题是沪深股市是否存在周效应，而数据文件中的交易日期是 yy/mm/dd 格式的日期型变量，因此首先要求出各日期所对应的星期数，这就要用到 SPSS

转换模块中的计算变量。

在 SPSS 中进行变量计算，可以利用“ 转换→计算变量 ”命令实现，菜单内容如图 18.1 所示。



图 18.1　SPSS 计算变量模块

计算变量的 SPSS 操作如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“ 转换→计算变量 ”，打开“ 计算变量 ”对话框，如图 18.2 所示。



图 18.2 “ 计算变量 ”对话框

在对话框的“ 目标变量 ”框中输入“ 星期 ”，定义生成的新变量；在“ 函数组 ”列表框中，选择“ 抽取日期 ”，在下面的“ 函数和特殊变量 ”列表框中选择“ XDATA.WKDAY ”，单击按钮，其选入到“ 数字表达式 ”列表框中；单击选中变量“ 交易日期[交易日期] ”，单击按钮，将其移动到右侧的“ 数字表达式 ”列表框中。如图 18.3 所示。



图 18.3　选入变量后的对话框

单击图 18.2 计算变量对话框“ 类型与标签 ”按钮，打开“ 计算变量：类型和标签 ”对话框，如图 18.4 所示，在“ 标签 ”编辑框中输入“ 星期 ”，单击“ 继续 ”按钮，关闭该对话框。

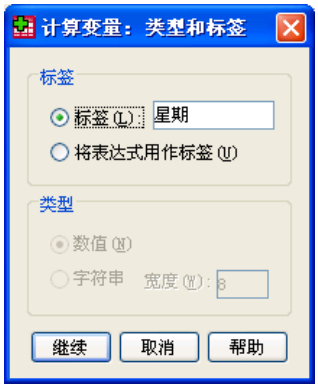


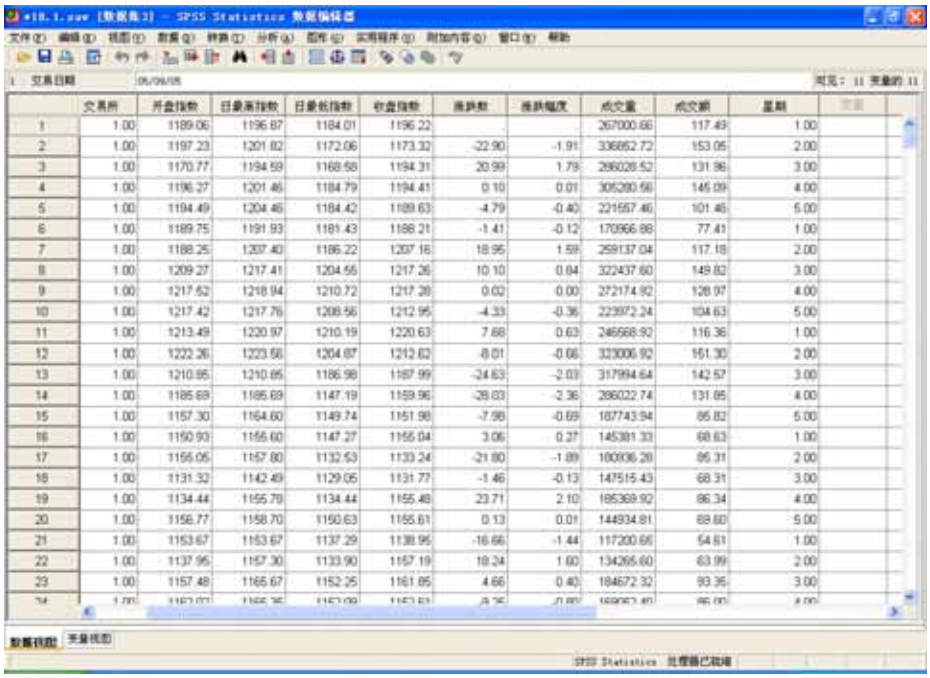
图 18.4 “ 计算变量：类型和标签 ”对话框

03 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 3.
COMPUTE 星期=XDATE.WKDAY(交易日期)-1.
VARIABLE LABELS  星期 '星期'.
EXECUTE.
  /SAVE=CLASS SCORES PROBS
```

04 结果输出。

设置完毕后，单击“确定”按钮，执行计算变量。得到的结果如图 18.5 所示，最后一列为生成的变量“星期”，它代表了各交易日所对应的星期数。



	交易所	开盘指数	日最高指数	日最低指数	收盘指数	涨跌数	涨跌幅度	成交量	成交额	星期
1	1.00	1189.06	1196.87	1184.01	1196.22			267000.66	117.49	1.00
2	1.00	1197.23	1201.82	1172.06	1173.32	-22.90	-1.91	336852.72	153.05	2.00
3	1.00	1170.77	1194.59	1168.58	1194.31	20.99	1.79	286028.52	131.96	3.00
4	1.00	1196.27	1201.46	1184.79	1194.41	0.10	0.01	305280.56	145.09	4.00
5	1.00	1194.49	1204.46	1184.42	1189.63	-4.79	-0.40	221557.46	101.46	5.00
6	1.00	1189.75	1191.93	1181.43	1188.21	-1.41	-0.12	170966.88	77.41	1.00
7	1.00	1188.25	1207.40	1186.22	1207.16	18.95	1.59	259137.04	117.18	2.00
8	1.00	1209.27	1217.41	1204.55	1217.26	10.10	0.84	322437.60	149.82	3.00
9	1.00	1217.62	1218.94	1210.72	1217.28	0.02	0.00	272174.92	128.97	4.00
10	1.00	1217.42	1217.76	1208.56	1212.95	-4.33	-0.36	223972.24	104.63	5.00
11	1.00	1213.49	1220.97	1210.19	1220.63	7.68	0.63	246568.92	116.36	1.00
12	1.00	1222.26	1223.56	1204.87	1212.62	-8.01	-0.66	323006.92	151.30	2.00
13	1.00	1210.85	1210.85	1186.98	1187.99	-24.63	-2.03	317994.64	142.57	3.00
14	1.00	1185.69	1185.69	1147.19	1159.96	-28.03	-2.36	286022.74	131.85	4.00
15	1.00	1157.30	1164.60	1149.74	1151.98	-7.98	-0.69	187743.94	86.82	5.00
16	1.00	1150.93	1155.60	1147.27	1155.04	3.06	0.27	145381.33	68.63	1.00
17	1.00	1156.05	1157.80	1132.53	1133.24	-21.80	-1.89	180936.28	85.31	2.00
18	1.00	1131.32	1142.49	1129.05	1131.77	-1.46	-0.13	147515.43	68.31	3.00
19	1.00	1134.44	1155.79	1134.44	1155.48	23.71	2.10	185369.92	86.34	4.00
20	1.00	1156.77	1158.70	1150.63	1155.61	0.13	0.01	144934.81	68.60	5.00
21	1.00	1153.67	1153.67	1137.29	1138.95	-16.66	-1.44	117200.66	54.61	1.00
22	1.00	1137.96	1157.30	1133.90	1157.19	18.24	1.60	134265.60	63.99	2.00
23	1.00	1157.48	1165.67	1152.25	1161.85	4.66	0.40	184672.32	93.35	3.00

图 18.5　计算变量输出结果

需要注意的是，函数 XDATA.WKDAY()可以将日期转化为其对应的星期数，但是在国外由于通常把星期日作为一周的第一天，因此，SPSS 的转换标准时将星期日赋值为 1，星期一赋值为 2……依次类推，为了和国内通常的星期表达方式一致，这里在做转换的时候，将函数 XDATA.WKDAY()所得的值全部减 1。对于星期日，SPSS 下取值为 1，减 1 之后变为 0。但是因为我国证券市场只在星期一到星期五进行交易，所以这种特殊的情况可以不予考虑。

18.1.3.2　描述性统计量

选取沪深股市的指数每日涨跌幅度和成交量两项指标来研究其周效应。

在 SPSS 中，计算描述性统计量可以利用“分析→描述统计→频数”命令实现，菜单

内容如图 18.6 所示。



图 18.6　SPSS 频率分析模块

频率分析的 SPSS 操作如下：

01 建立数据文件。

利用上一步得到的结果数据。

02 拆分数据文件。

为了进行频率分析，首先应拆分文件。在 SPSS 中，可以利用“ 数据→拆分文件 ”命令实现，菜单内容如图 18.7 所示。

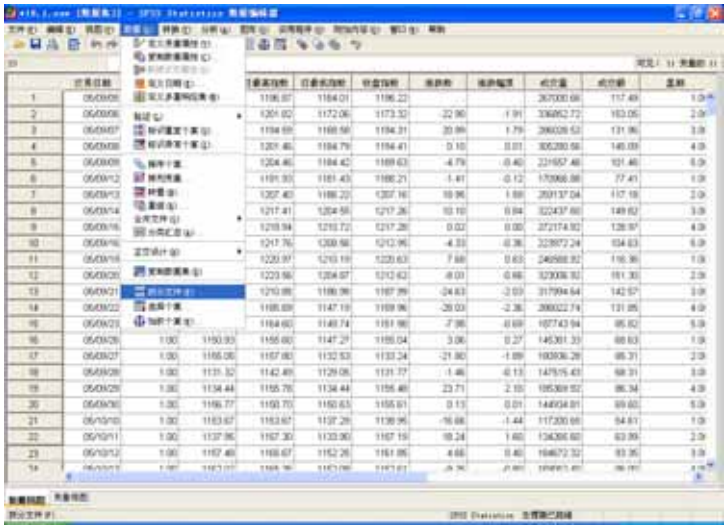


图 18.7　SPSS 拆分文件模块

依次单击菜单“ 数据→拆分文件 ”，打开“ 分割文件 ”对话框，如图 18.8 所示。

在图 18.8 所示对话框中，选择“ 比较组 ”，再将左侧变量列表框中的变量“ 交易所[交易所] ”和“ 星期[星期] ”移至右侧的“ 分组方式 ”列表框中，如图 18.9 所示。



图 18.8　“ 分割文件 ”对话框

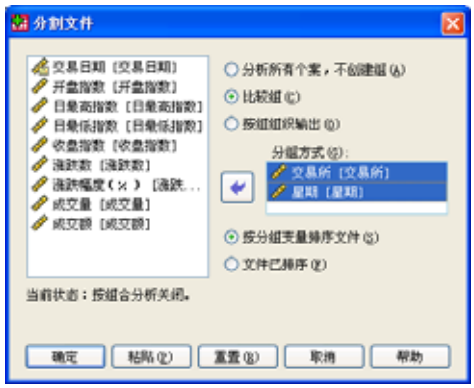


图 18.9　变量选入后的“ 分割文件 ”对话框

单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
SORT CASES  BY 交易所 星期.
SPLIT FILE LAYERED BY 交易所 星期.
```

单击“ 确定 ”按钮，执行数据拆分。

03 频率分析选入变量及参数设置。

依次单击菜单“ 分析→描述统计→频率 ”命令，打开“ 频率 ”对话框。如图 18.10 所示。

选择对话框的左侧变量列表框中的变量“ 涨跌幅度（%）[涨跌幅度（%）] ”和“ 成交量[成交量] ”，单击按钮，将其选入到右侧的“ 变量 ”列表框。如图 18.11 所示。



图 18.10　“ 频率 ”对话框



图 18.11　变量选入后的“ 频率 ”对话框

- 频率分析：统计量对话框

单击图 18.10 中“ 统计量 ”按钮，打开“ 频率：统计量 ”对话框。如图 18.12 所示。该对话框包含“ 百分位值 ”、“ 集中趋势 ”、“ 离散 ”和“ 分布 ”4 个选项组。

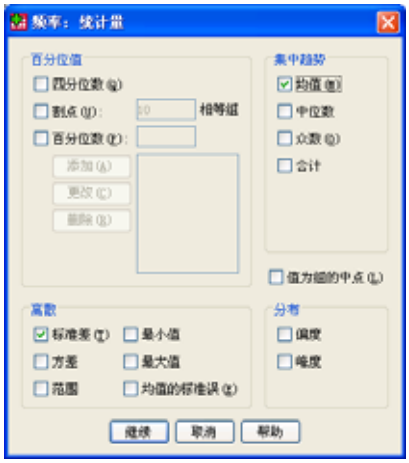


图 18.12 “ 频率：统计量 ” 对话框

在 “ 百分位值 ” 选项组中，各选项及其含义如表 18.2 所示。

表 18.2 百分位值各选项及其含义

选 项 名	含 义
四分位数	输出四分位数，显示 25%、50%、75%的百分位数
割点	将数据平均分为所设定的相等等份，可输入 2～100 之间的任意整数，如键入 4，则输出第 25、50、75 百分位数
百分位数	自定义百分位数，可输入 0～100 之间的任意整数。输入值后，按 “ 添加 ” 键可以将值添加到右侧列表框，可反复操作键入多个百分位数；按 “ 删除 ” 键则删除已键入的数值；按 “ 更改 ” 键则重新输入新数

“ 离散 ”选项组主要用于设置输出结果的离差，包括“ 标准差 ”、“ 方差 ”、“ 范围 ”、“ 最大值 ”、“ 最小值 ” 和 “ 均值的标准误 ” 等。

“ 集中趋势 ”选项组主要用于设置输出结果的集中趋势，包括“ 均值 ”、“ 中位数 ”、“ 众数 ” 和 “ 合计 ”。

“ 分布 ” 选项组主要用于设置输出结果的分布参数，包括 “ 偏度 ” 和 “ 峰度 ” 两项。

“ 值为组的中点 ” 表示在计算百分位数值和中位数时，假设数据已分组，且用各组的组中值代替各组数据。

本例在 “ 离散 ” 选项组选择 “ 标准差 ”，“ 集中趋势 ” 选项组选择 “ 均值 ”，其他选项组参数选择默认。

- 频率分析：图表对话框

单击图 18.10 中 “ 图表 ” 按钮，打开 “ 频率：图表 ” 对话框。如图 18.13 所示。

在该对话框中，可以根据资料的情况选择正确的图形。区间（尺度）数据应该选择直方图，而非区间数据可以选择其他图形。



图 18.13 “ 频率：图表 ” 对话框

“ 图表类型 ” 选项组各选项及其含义如表 18.3 所示。

表 18.3　图表类型各选项及其含义

选　项	意　义
无	不输出任何图形，为默认
条形图	输出条形图，各条高度代表变量各分类的频数
饼图	输出饼图，饼图中各块的数值表示各分类变量的频数
直方图	输出直方图，此图仅适用于区间型数值变量。选择此项后，还可选择 “ 带正态曲线 ”，使画出的直方图带有正态曲线

“ 图表值 ” 选项组只有选择了条形图和饼图项才有效，决定纵轴表示的统计量。各选项含义如表 18.4 所示。

表 18.4　图表值各选项及其含义

选　项	含　义
频率	纵轴表示频率
百分比	纵轴表示百分比

本例中，各选项组参数均选择默认。

- 频率分析：格式对话框

单击图 18.10 中 “ 格式 ” 按钮，打开 “ 频率：格式 ” 对话框。如图 18.14 所示。

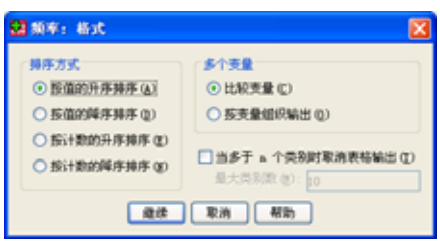


图 18.14 “ 频率：格式 ” 对话框

在该对话框中，主要包括“ 排序方式 ”和“ 多个变量 ”选项组。“ 排序方式 ”选项组用于确定频数表中的排列顺序；“ 多个变量 ”选项组用于设定多变量表格输出的格式；“ 当多于 n 个类别时取消表格输出 ”用于控制频数表输出的分类数量，默认值为 10。

04 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
FREQUENCIES VARIABLES=涨跌幅度 成交量
  /STATISTICS=STDDEV
  /ORDER=ANALYSIS.
```

05 频率分析结果输出。

设置完毕后，单击图 18.10 对话框中的“ 确定 ”按钮，执行频率分析。

- 沪深股市周效应分析表

表 18.5　沪深股市周效应分析表

统 计 量			
交易所　星期		涨跌幅度 (%)	成 交 量
沪市　1,00　N	有效	60	61
	缺失	7	0
	均值	.4213	337240.6918
	标准差	1．31157	1.83844E5
2.00　N	有效	61	61
	缺失	0	0
	均值	.1089	343049.4716
	标准差	1.18174	1.77644E5
3.00　N	有效	61	61
	缺失	0	0
	均值	.1910	332493.0916
	标准差	1.19435	1.51707E5
4.00　N	有效	60	60
	缺失	0	0
	均值	.0262	339727.6873
	标准差	1.24009	1.65527E5
5.00　N	有效	60	60
	缺失	O	0
	均值	.2805	342171.4563
	标准差	1.17372	1.77799E5
深市　1.00　N	有效	60	61
	缺失	1	0
	均值	.4705	42190.2689
	标准差	1.60792	24887.43033
2.00　N	有效	61	61
	缺失	0	0

续表

统 计 量			
交易所　星期		涨跌幅度(%)	成 交 量
	均值	.1426	43540.9938
	标准差	1.30311	27784.64210
3.00	N　有效	61	61
	缺失	0	0
	均值	.1793	42425.8954
	标准差	1.45930	24264.90523
4.00	N　有效	60	60
	缺失	0	0
	均值	.0788	42780.4783
	标准差	1.35941	23397.27055
5.00	N　有效	60	60
	缺失	0	0
	均值	.3002	41515.5915
	标准差	1.17326	421897.31461

表 18.5 为沪深两市一周的统计量信息，包括涨跌幅度和成交量的有效个案数、缺失个案数和标准差等。

18.1.3.3　图形绘制

将表格中的数据用图形直观表示出来。

在 SPSS 中，绘制图形可以利用“ 图形→旧对话框→条形图 ”命令实现，菜单内容如图 18.15 所示。



图 18.15　SPSS 条形图绘制模块

条形图绘制的 SPSS 操作如下：

01 建立数据文件。

利用上一步得到的结果数据。

02 选入变量及参数设置。

依次单击菜单“ 图形→旧对话框→条形图 ”命令，打开“ 条形图 ”对话框，如图 18.16 所示。



图 18.16 “ 条形图 ”对话框

在定义选项框的上方有 3 种条形图可选：简单条形图、复式条形图、堆积面积图。本例选择复式条形图。

- 条形图：数据类型选项组

在图 18.16 所示的对话框中，“ 图表中的数据为 ”选项组可以定义数据类型，系统共提供了 3 种数据类型，如表 18.6 所示。

表 18.6 数据类型选项及其含义

选　项	含　义
个案组摘要	以组为单位体现数据
各个变量的摘要	以变量为单位体现数据
个案值	以观察样例为单位体现数据

大多数情况下，统计图都是以组为单位的形式来体现数据的。

- 定义复式条形图：个案组摘要对话框

单击图 18.16 中“ 定义 ”按钮，打开“ 定义复式条形图：个案组摘要 ”对话框。如图 18.17 所示。

首先，在对话框的“ 条的表征 ”选项组选择“ 其他统计量（例如均值）”选项，激活“ 变量 ”列表框；然后，在左侧变量列表框中选择变量“ 涨跌幅度（%）[涨跌幅度（%）] ”，

单击按钮，将其移到右侧的“ 变量 ”列表框。将变量“ 星期［星期］”选入“ 类别轴 ”列表框，将变量“ 交易所 ”选入“ 定义聚类 ”列表框，如图 18.18 所示。



图 18.17　“ 定义复式条形图：个案组摘要 ”对话框



图 18.18　变量选入后的“ 定义复条形图：个案组摘要 ”对话框

单击“ 更改统计量 ”按钮，打开“ 统计量 ”对话框，如图 18.19 所示。

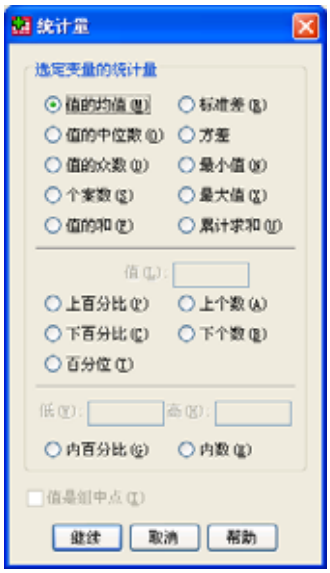


图 18.19　“ 统计量 ”对话框

在该对话框“选定变量的统计量”选项组中，选择“值的均值”；其他选项组按默认选项。

- 定义复式条形图：选项对话框

单击图 18.17 中“选项”按钮，打开“选项”对话框。如图 18.20 所示。

该对话框用于设置缺失值处理方式、误差条形图显示细节等。

03 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 4.
GRAPH
  /BAR(GROUPED)=MEAN(涨跌幅度) BY 星期 BY 交易所.
```

04 结果输出。

设置完毕后，单击“确定”按钮，执行复式条形图绘制。

- 均值涨跌幅度-星期复式条形图



图 18.20 “选项”对话框

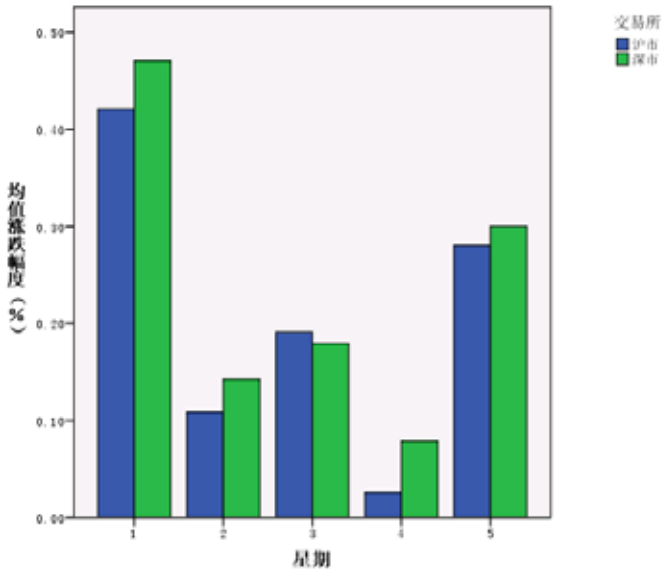


图 18.21 “均值涨跌幅度-星期”复式条形图

图 18.21 为“均值涨跌幅度-星期”复式条形图。从图中可见，沪深两市星期一的涨跌幅度最大，星期四的涨跌幅度最小。

- 标准差涨跌幅度-星期复式条形图

在图 18.19 所示对话框中的“选定变量的统计量”选项组中，将“值的均值”改为“标准差”，执行得到“标准差涨跌幅度-星期”复式条形图，如图 18.22 所示。

图 18.22 为“标准差涨跌幅度-星期”复式条形图。从图中可以看出，沪深两市的涨跌幅度之间没有显著差异。

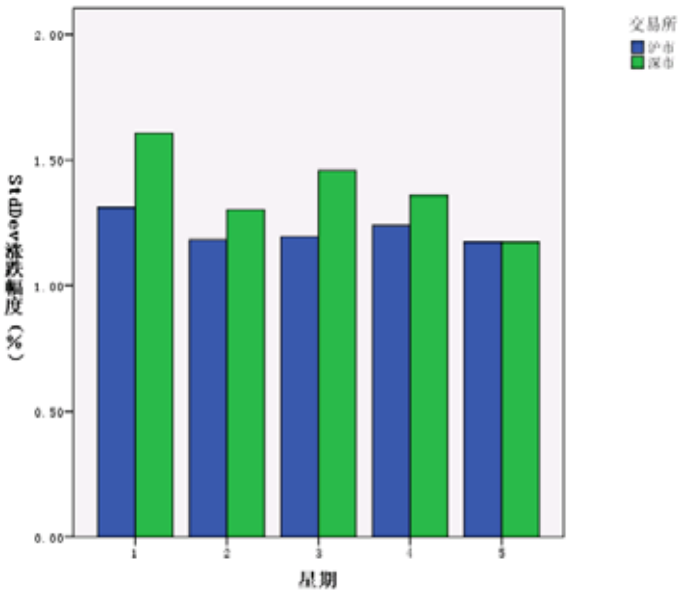


图 18.22 “ 标准差涨跌幅度-星期 ” 复式条形图

- 均值成交量-星期复式条形图

在图 18.18 对话框中，将 “ 成交量[成交量] ” 选入 “ 变量 ” 列表框；在图 18.19 所示对话框中的 “ 选定变量的统计量 ” 选项组中选择 “ 值的均值 ”，执行得到 “ 均值成交量-星期 ” 复式条形图，如图 18.23 所示。

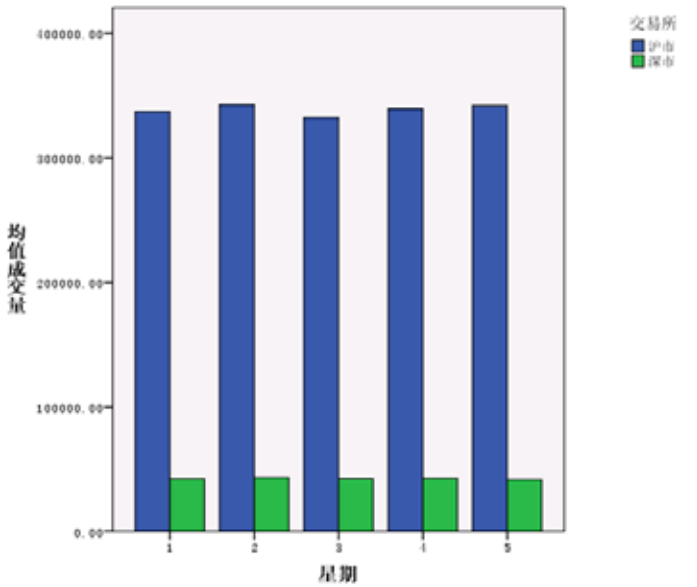


图 18.23 “ 均值成交量-星期 ” 复式条形图

图 18.23 为“ 均值成交量-星期 ”复式条形图，从图中可以看出，沪深两市的成交量在一周各日内没有显著差异。

- 标准差成交量-星期复式条形图

在图 18.18 所示对话框中，将“ 成交量[成交量] ”选入“ 变量 ”列表框；将图 18.19 对话框中的“ 选定变量的统计量 ”选项组中的“ 值的均值 ”改为“ 标准差 ”，执行得到“ 标准差成交量-星期 ”复式条形图，如图 18.24 所示。

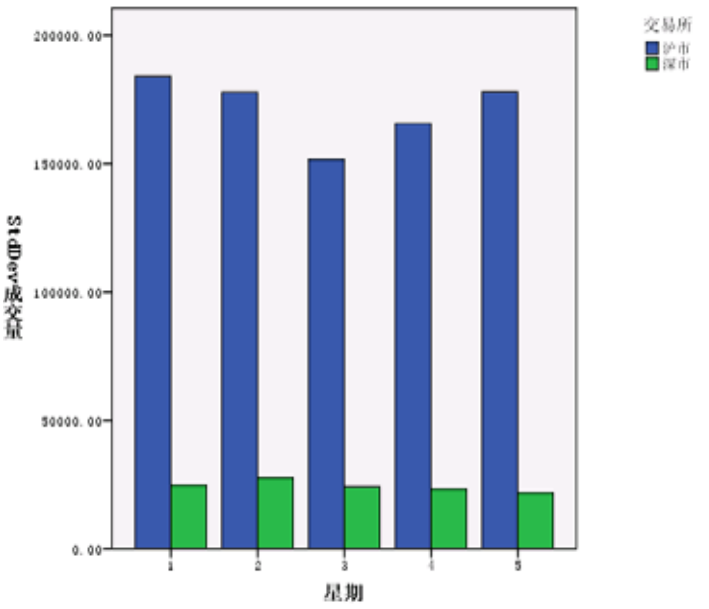


图 18.24 “ 标准差成交量-星期 ”复式条形图

图 18.24 为“ 标准差成交量-星期 ”复式条形图，从图中可以看出，沪深两市的成交量表现出明显的“ 沪强深弱 ”的趋势。

18.1.3.4 对模型的进一步检验—单因素方差分析

前面根据 SPSS 所计算的描述性统计量及绘制的图形给出了 4 条结论，但是这只是一种主观的判断，有必要对这些结论进行进一步的检验。首先分别对沪深两市的涨跌幅度做单因素方差分析，考察我国股市是否真的存在周效应。

在 SPSS 中，执行单因素方差分析可以利用“ 分析→比较均值→单因素 ANOVA ”命令实现，菜单内容如图 18.25 所示。

单因素方差分析的 SPSS 操作如下：

01 建立数据文件。

将计算变量后的数据文件导入到 SPSS 中。

02 拆分数据文件。

为了进行单因素方差分析，首先应拆分文件。在 SPSS 中，可以利用“ 数据→拆分文

件 ” 命令实现，菜单内容如图 18.26 所示。

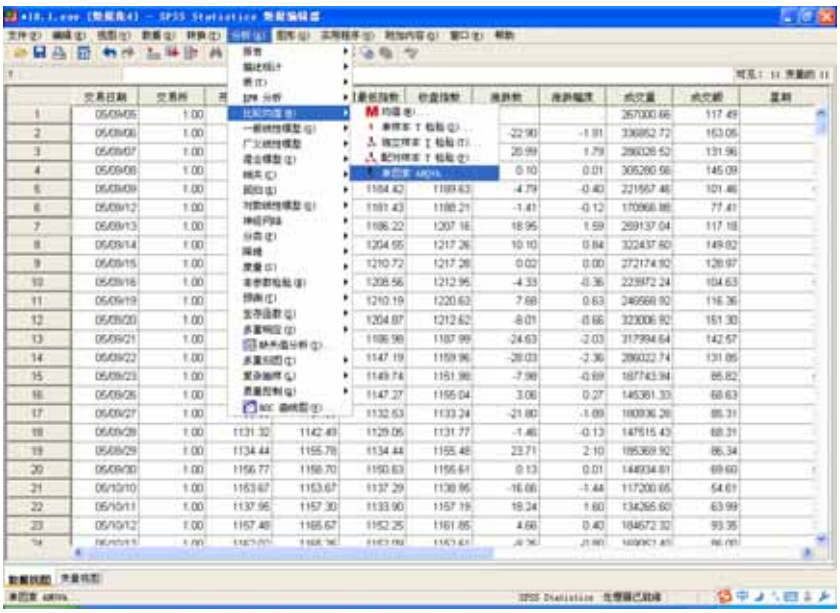


图 18.25　SPSS 单因素方差分析模块

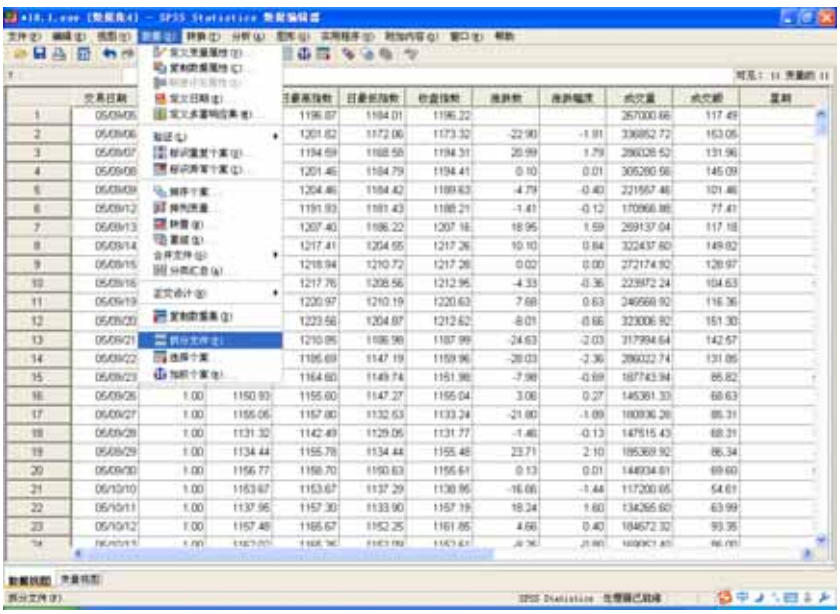


图 18.26　SPSS 拆分文件模块

依次单击菜单 “ 数据→拆分文件 ” 命令，打开 “ 分割文件 ” 对话框，如图 18.27 所示。在图 18.27 所示对话框中，选择 “ 比较组 ”，再将左侧变量列表框中的 “ 交易所[交易

所]”变量移至右侧“分组方式”列表框中，如图 18.28 所示。



图 18.27 “分割文件”对话框

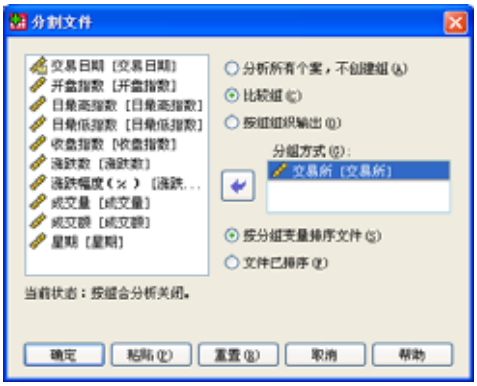


图 18.28 变量选入后的“分割文件”对话框

单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
SORT CASES  BY 交易所.
SPLIT FILE LAYERED BY 交易所.
```

单击“确定”按钮，执行数据拆分。

03 单因素方差分析选入变量及参数设置。

依次单击菜单“分析→比较均值→单因素 ANOVA”，打开“单因素方差分析”对话框。如图 18.29 所示。



图 18.29 “单因素方差分析”对话框

在对话框的左侧变量列表框中选择变量“涨跌幅度（%）[涨跌幅度（%）]”，单击按钮，将其选入到右侧的“因变量”列表框；用同样方法将“星期[星期]”选入到“因子”列表框。如图 18.30 所示。

- 单因素方差分析：对比对话框

单击图 18.29 中的“对比”按钮，打开“单因素 ANOVA：对比”对话框，如图 18.31 所示。





图 18.30　变量选入后的“ 单因素方差分析 ”对话框　图 18.31　“ 单因素 ANOVA：对比 ”对话框

该对话框有两个选项组：“ 多项式 ”和“ 1 的对比 1 ”。

“ 多项式 ”选项用于将组间平方划分为趋势成分，可对因变量按因子变量中的水平次序进行趋势性检验。“ 度 ”下拉菜单用于设定多项式的次数，可选项有：线性多项式、二次项式、立方、四次项式和五次项式。

“ 1 的对比 1 ”选项区域中的“ 系数 ”需要用户指定检验对比将通过 T 检验，可以输入因变量中每组的系数，系数的顺序对应于因子变量中分类值的递增顺序，列表中的第一系数对应于分类变量的最小值，最后一个系数对应于最大值。

- 单因素方差分析：两两比较对话框

单击图 18.29 中“ 两两比较 ”按钮，打开“ 单因素 ANOVA：两两比较 ”对话框，如图 18.32 所示。



图 18.32　“ 单因素 ANOVA：两两比较 ”对话框

在该对话框中进行均值的多重比较，各选项的含义如下所述。

“ 假定方差齐性 ”选项组给出当方差相等时的确定多重比较方法，包括 14 种可选方法，

具体如表 18.7 所示。

表 18.7　假定方差齐性各选项及其含义

方 法 名	含　义
LSD	Least-significant difference，即最小显著性差异法。用 T 检验完成各组间的配对比较，检验的敏感性高，各个水平间的均值存在微小的差异也有可能被检验出来，但此方法对第一类弃真错误的概率不进行控制和调整
Bonferron i	LSDMOD 方法，即修正最小显著差异法。用 T 检验完成各组间的配对比较；同时，通过设置每个检验的误差率来控制第一类错误的概率
Sidak	Sidak 法，根据 *t* 统计量进行配对多重比较，调整多重比较的显著性水平，其界限比 Bonferroni 方法小
Scheffe	Scheffe 差别检验法。α 可以指定 0~1 之间任何显著性水平，默认值为 0.05
R-E-G-W-F	Ryan-Einot-Gabriel-Welsch F 方法，根据 *F* 统计量检验的 Ryan-Einot-Welsch 多重比较
R-E-G-W-Q	Ryan-Einot-Gabriel-Welsch Q 方法，根据 Student 极差统计量的 Ryan-Einot-Welsch 多重比较
S-N-K	Student-Newman-Keuls 方法，根据 Student 极差统计量在均值间进行配对比较；如果各组样本含量相等或选择了“ Harmonic average of all groups ”(所有个组样本含量的调和平均值)，即用所有各组样本含量的调和平均值进行样本量的估计时，还用逐步过程进行齐次子集(差异较小的子集)的均值配对比较；在该比较过程中各组均值按从大到小的顺序排列，最先比较最末端的差异
Tukey	Tukey’s honestly significant difference 方法，即 Tukey 显著差异法，用 Student-Range 统计量进行所有组间均值的配对比较，用所有配对比较误差率作为实验误差率
Tukey’s-b	Tukey-b 方法，用 Student-Range 分布进行组间均值的配对比较，其精确值为两种检验相应值的平均值
Duncan	Duncan 方法，指定一系列的 Range 值逐步进行计算比较得到结论
Hochberg’s GT2	Hochberg’s GT2 方法，用正态最大系数进行多重比较
Gabriel	Gabriel 方法，用正态标准系数进行配对比较，单元较大时这种方法较自由
Waller-Duncan	Waller-Duncan 方法，用 t 统计量进行多重比较检验，使用贝叶斯过程的多重比较检验
Dunnett	Dunnett 方法，指定此方法进行配对比较，方法是选择最后一组为对比，其他各组和它比较，选定此方法后将激活“ Control Catetory ”栏，在该栏展开下拉菜单选择对比组，其中可选“ Last ”分类和“ First ”分类；并在“ Test ”选项组中设定单双侧检验的选择，其中可选“ 2-Sided、<Control、>Control ”

“ 未假定方差齐性 ”选项组中给出当方差不齐时的多重比较方法选项，其含义如表 18.8 所示。

表 18.8　未假定方差齐性各选项及其含义

方 法 名	含　义
Tamhane’s T2	用 T 检验进行配对比较试验
Dunnett’s T3	用基于 Student 最大模数的比较配对试验
Games Howell	方差不齐时的配对比较试验
Dunnett’s C	用基于 Student 极差的配对比较试验

本例中，在“ 假定方差齐性 ”选项组选择“ LSD ”。

● 单因素方差分析：选项对话框

单击图 18.29 中的“ 选项 ”按钮，打开“ 单因素 ANOVA：选项 ”对话框。如图 18.33 所示。



图 18.33 “ 单因素 ANOVA：选项 ”对话框

该对话框给出了检验统计量、检验和缺失值的处理方式等设置。

“ 统计量 ”选项组为输出统计量的选项，其选项及其含义如表 18.9 所示。

表 18.9　统计量各选项及其含义

选 项 名	含　义
描述性	要求输出描述统计量，其中包括观测量数目均值、标准差、标准误差、最小值、最大值以及各组中每个因变量的 95%置信区间
固定和随机效果	显示固定效应模型的标准差、标准误差和 95%的置信区间；随机效应模型的标准误差、95%的置信区间及方差成分间的估计值
方差同质性检验	要求计算分组树木相等的 Levene 统计量，该检验方法不依赖于正态假设
Brown-Forsythe	计算分组均数相等的 Brown-Forsythe 统计量。当不能把握方差齐性假设时，此统计量比 F 统计量更有优势
Welch	计算分组均数相等的 Welch 统计量。当不能把握方差齐性假设时，此统计量比 F 统计量更有优势

该对话框给出“ 均值图 ”选项，选择该项，表示显示分组均值图。

“ 缺失值 ”选项组用于选择缺失值处理方法，选项及其含义如表 18.10 所示。

表 18.10　缺失值处理方法

处理方法	含　义
按分析顺序排除个案	剔除各分析中含有缺失值的个数
按列表排除个案	剔除含缺失值的全部个案

本例中，“ 统计量 ” 选项组选择 “ 方差同质性检验 ”；绘制均值图。

04 单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
ONEWAY 涨跌幅度 BY 星期
  /STATISTICS HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=LSD ALPHA(0.05).
```

05 涨跌幅度单因素方差分析结果输出

设置完毕后，单击图 18.29 对话框中的 “ 确定 ” 按钮，执行单因素方差分析。

- 涨跌幅度方差齐性检验

表 18.11　涨跌幅度方差齐性检验

方差齐性检验

涨跌幅度

交 易 所	Levene 统计量	Df1	Df2	显 著 性
沪市	.337	4	297	.853
深市	1.209	4	297	.307

表 18.11 为涨跌幅度方差齐性检验结果。从表中可以看出，沪深两市各星期间涨跌幅度的 Levene 统计量分别为 0.337 和 1.209，第一自由度均为 4，第二自由度均为 297，显著性检验分别为 0.853 和 0.307，大于 0.005，所以，可以认为方差是齐性的。

- 涨跌幅度方差分析表

表 18.12　涨跌幅度方差分析表

ANOVA 涨跌幅度

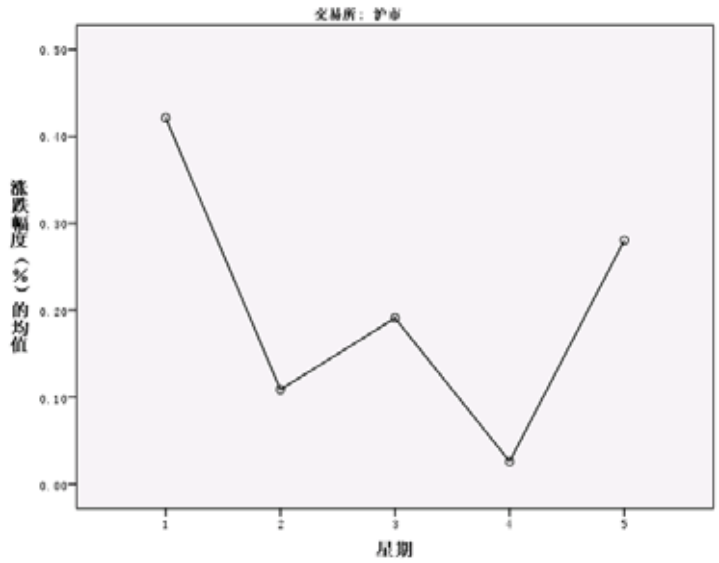
交 易 所		平 方 和	df	均　方	F	显 著 性
沪市	蛆间	5.645	4	1.411	.946	.437
	组内	442.883	297	1.491		
	总数	448.528	301			
深市	组间	5.755	4	1.439	.746	.561
	蛆内	572.446	297	1.927		
	总数	578.200	301			

表 18.12 为涨跌幅度方差分析表。从表中可以看出，第五列是统计量 F 值，“ 组间 ” 和 “ 组内 ” 为方差的两个来源，第三列是检验统计量的自由度（df），沪市和深市的组间自由度均为 4，组内自由度均为 297，总自由度也相同，为 301。由此可以看出，虽然前面的描述性统计量分析认为星期一涨跌幅度的均值大于其他各日，但是这种差异是不显著的。

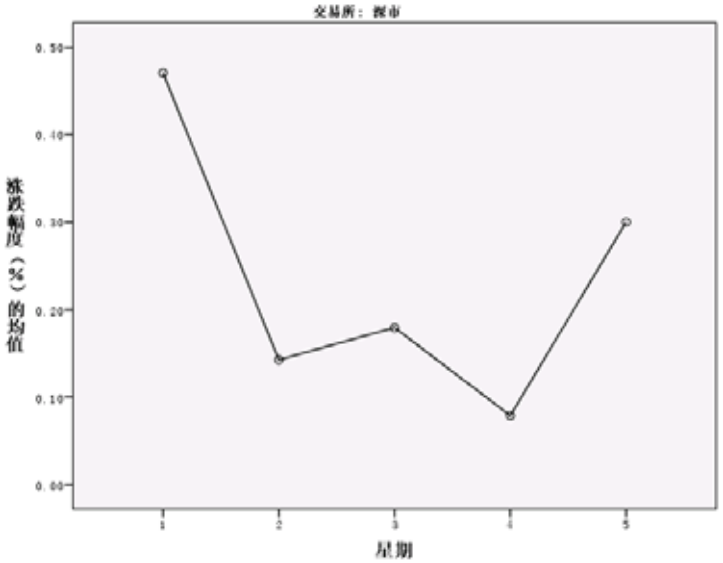
- 涨跌幅度均值图

图 18.34 是沪深两市股票涨跌幅度均值图。结合图形和表 18.12 的方差分析表可以判断，虽然股票在一星期的各日之内涨跌幅度有所差异，但是这种差异是不显著的，不能因

此就片面地判断我国的证券市场的指数涨跌幅度存在周效应。



（a）沪市涨跌幅度均值图



（b）深市涨跌幅度均值图

图 18.34　涨跌幅度均值图

06 交易量单因素方差分析结果输出

在图 18.29 所示对话框的左侧变量列表框中选择变量“ 交易量[交易量] ”，单击 按钮，将其选入到右侧的“ 因变量 ”列表框；用同样方法将“ 星期[星期] ”选入到“ 因子 ”列表框。如图 18.35 所示。



图 18.35　变量选入后的“ 单因素方差分析 ”对话框

其他参数设置和“ 涨跌幅度单因素方差分析 ”相同，单击“ 确定 ”按钮，对交易量做单因素方差分析。

- 交易量方差分析表

表 18.13　交易量方差分析表

ANOVA

成交量

交 易 所		平 方 和	df	均　方	F	显 著 性
沪市	组间	4.405E9	4	1.10PE9	.037	.997
	组内	8.784E12	298	2．948E10		
	总数	8.788E12	302			
繡市	组间	1.351E8	4	3.378E7	.056	.994
	组内	1.794E11	298	6.020E8		
	总数	1 795E11	302			

表 18.13 为交易量方差分析表。沪市和深市的组间自由度均为 4，组内自由度均为 298，总自由度也相同，均为 302。

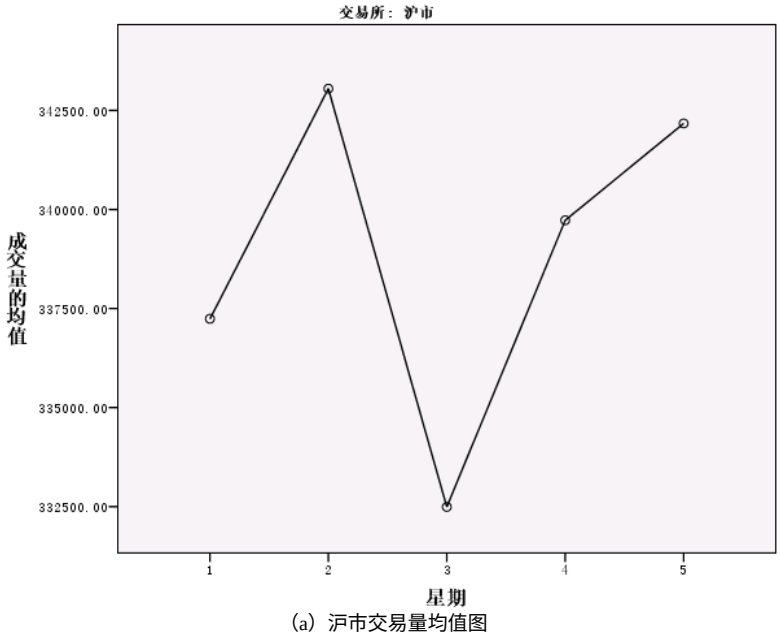
- 交易量均值图

图 18.36 是沪深两市股票交易量均值图。结合该图和表 18.13 的方差分析表可以判断，沪深两市在一周内的日成交量虽然有一定差异，但是这种差异是不显著的，不能认为我国的证券市场的成交量存在周效应。

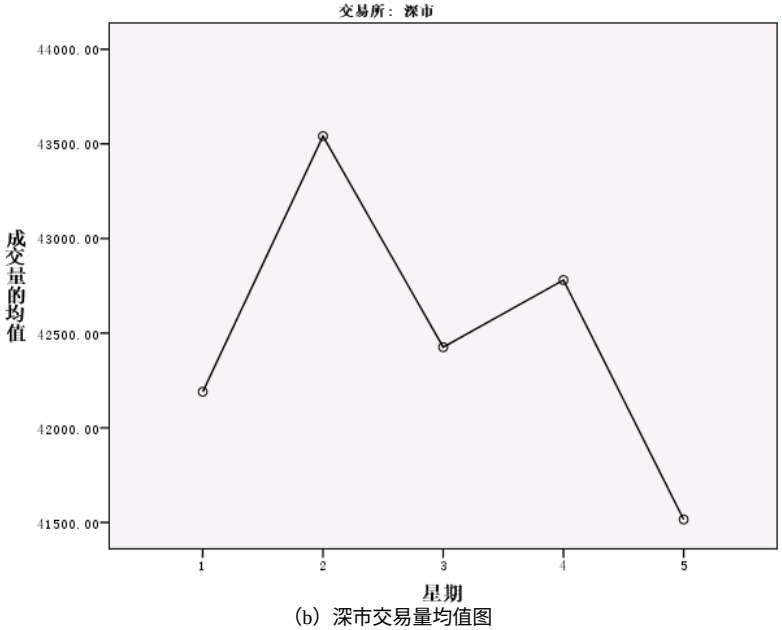
18.1.3.5　对模型的进一步检验—相关分析

从图 18.34 和图 18.36 可以看出，沪深两市的涨跌幅度在一周内的均值曲线以星期三为中心呈现 W 型，而日成交量的均值曲线正好近似以星期三为中心呈现出 M 型，因此可以用相关分析寻找二者之间的相关性。

在 SPSS 中，执行双变量相关分析可以利用“ 分析→相关→双变量 ”命令实现，菜单内容如图 18.37 所示。



(a) 沪市交易量均值图



(b) 深市交易量均值图

图 18.36　交易量均值图

双变量相关分析的 SPSS 操作如下：

01 建立数据文件。

将计算变量后的数据文件导入到 SPSS 中。

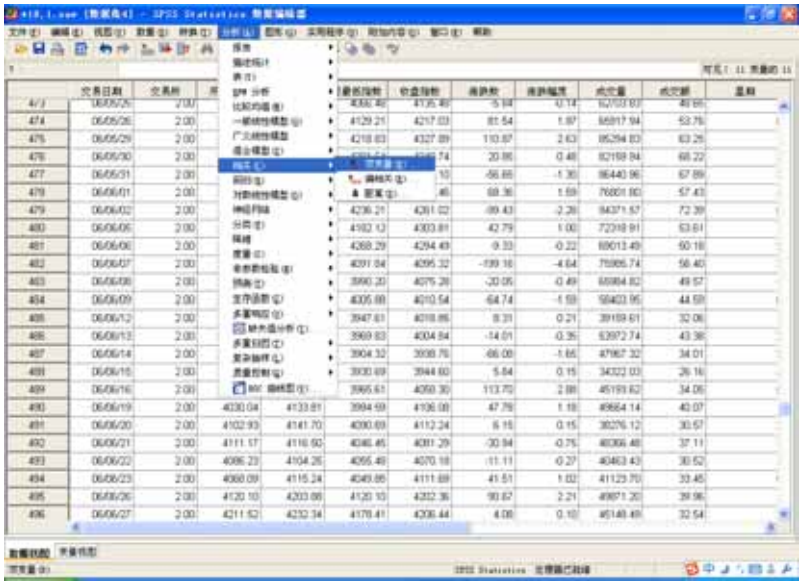


图 18.37　SPSS 双变量相关分析模块

02 双变量相关分析选入变量及参数设置。

依次单击菜单“分析→相关→双变量”命令，打开“双变量相关”对话框，如图 18.38 所示。



图 18.38 “双变量相关”对话框

在对话框的左侧变量列表框中选择变量“涨跌幅度（%）[涨跌幅度（%）]”和“成交量[成交量]”，单击 按钮，将其选入到右侧的“变量”列表框，如图 18.39 所示。



图 18.39　变量选入后的“ 双变量相关 ”对话框

● 双变量相关分析：相关系数设置

在图 18.38 中的“ 相关系数 ”选项组中可以选择相关系数的类型，共有 3 种，如表 18.14 所示。

表 18.14　相关系数类型及其含义

类 型 名	含　义
Pearson	相关系数，SPSS 默认选项
Kendell 的 tau-b	非参数资料的相关系数
Spearman	非正态分布资料的 Pearson 相关系数替代值

本例选择 Pearson 和 Spearman。

● 双变量相关分析：显著性检验设置

在图 18.38 中所示“ 显著性检验 ”选项组中可以相关系数的显著性检验类型，共有两种：相关系数的单侧（One-tailed）或双侧（Two-tailed）检验。本例选双侧检验。

● 双变量相关分析：选项对话框

单击图 18.38 中的“ 选项 ”按钮，打开“ 双变量相关性：选项 ”对话框，如图 18.40 所示。



图 18.40　“ 双变量相关性：选项 ”对话框

通过这个对话框，可以选择输出统计量和缺失值的处理方法。

本例在“统计量”选项组选择“均值和标准差”。

03 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
CORRELATIONS
  /VARIABLES=成交量 涨跌幅度
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE.
NONPAR CORR
  /VARIABLES=成交量 涨跌幅度
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

04 涨跌幅度单因素方差分析结果输出。

设置完毕后，单击图 18.38 对话框中的“确定”按钮，执行双变量相关分析。

- 描述性统计量

表 18.15　描述性统计量

描述性统计量

交易所		均值	标准差	N
沪市	成交量	338923.920	1.70589E5	303
	涨跌幅度（%）	.2052	1.22071	302
课市	成交量	42492.9070	24381.96753	303
	潞跌幅度（%）	.2338	1.38598	302

表 18.15 为描述性统计量。变量“成交量”和“涨跌幅度（%）”的个案数、均值与标准差。

- Pearson 相关系数

表 18.16　Pearson 相关系数

相关性

交易所			成交量	涨跌幅度（%）
沪市	成交量	Pearson 相关性	1.000	.111
		显著性（双侧）		.055
		N	303	302
	涨跌幅度（%）	Pearson 相关性	.111	1
		显著性（双侧）	.055	
		N	302	302
深市	成交量	Pearson 相关性	1	.142*
		显著性（双侧）		.013
		N	303	302
	涨跌幅度（%）	Pearson 相关性	.142*	1
		显著性（双侧）	.013	
		N	302	302

*.在 0.05 水平（双侧）上显著相关。

表 18.16 为 Pearson 相关系数的结果，从表中可以看出，对于沪市，成交量和涨跌幅度的相关系数为 0.111，其显著性（双侧）检验结果为 0.055；对于深市，成交量和涨跌幅度的相关系数为 0.142，其显著性（双侧）检验结果为 0.013。说明成交量与涨跌幅度之间存在相关关系，但还不能确定二者具体的相关程度。

- Spearman 相关系数

表 18.17　Spearman 相关系数

相关系数

交易所			成交量	涨跌幅度（%）
Spearman 的 rho	沪市　成交量	相关系数	1.000	.128*
		Sig.（双侧）	.	.027
		N	303	302
	涨跌幅度（%）	相关系数	.128*	1.000
		Sig.（双侧）	.027	.
		N	302	302
	深市　成交量	相关系数	1.000	.214**
		Sig.（双侧）	.	.000
		N	303	302
	涨跌幅度（%）	相关系数	.214**	1.000
		Sig.（双侧）	.000	.
		N	302	302

*.在置信度（双测）为 0.05 时，相关性是显著的。

**.在置信度（双测）为 0.01 时，相关性是显著的。

表 18.17 为 Spearman 相关系数的结果，从表中可以看出，对于沪市，成交量和涨跌幅度的相关系数为 0.128，其显著性（双侧）检验结果为 0.027；对于深市，成交量和涨跌幅度的相关系数为 0.214，其显著性（双侧）检验结果为 0.000。说明成交量与涨跌幅度之间存在相关关系，但是还不能确定二者具体的相关程度，结论和表 18.17 类似。

通过上述分析可以看出，沪深两市的涨跌幅度和成交量之间应该有一定的相关关系，但仍需进一步的分析。并且涨跌幅度大致呈 W 型，成交量大致呈 M 型。沪深两市在一周各天虽然涨跌幅度和成交量都各有差异，但是这种差异都不显著。因此，不能简单地判定我国的股票市场存在着明显的周效应。

18.2　汇率影响因素分析

18.2.1　实例内容说明

现有 1985～2000 年的有关数据，如表 18.18 所示，试分析人民币及其影响因素的相关

性。所用数据参考自“人民币汇率研究”（陈璋，CENET 网刊，2005）、“汇率决定模型与中国汇率总分析”（孙煜，复旦大学《经济学人》，2004）和“人民币汇率的影响因素与走势分析”（徐晨，对外经济贸易大学硕士论文，2002），其中通货膨胀率、一年期名义利率、美元利率和汇率 4 个指标的数据来自于《中国统计年鉴》（2001，中国统计出版社）；2000 年的部分数据来自于国家统计局官方网站。

表 18.18　1985～2000 年的汇率及其影响因素数据

年份	通货膨胀率	一年期名义利率	美元利率	汇率	GDP（亿元）	净出口（亿美元）	居民总储蓄（亿元）	居民消费（亿元）	外商直接投资（亿美元）	实使外资（亿美元）	外汇储备（亿美元）	外债规模（亿美元）
1985	8.8	7.2	8	2.94	8964.4	-149	3042.7	4625.7	63.33	19.56	26.44	158.3
1986	6	7.2	6.5	3.45	10202.2	-119.7	4163.36	5214.1	33.3	22.44	20.72	214.8
1987	7.3	8.64	7	3.72	11962.5	-37.7	4642.08	6011.5	37.09	23.14	29.23	302
1988	18.5	11.34	7.5	3.72	14928.3	-77.5	7099.94	7694.1	52.97	31.94	33.72	400
1989	17.8	10.08	9	3.76	16909.2	-66	6996.74	8588	56	33.93	55.5	413
1990	2.1	8.64	8	4.78	18547.9	87.4	11078.36	9108.9	65.96	34.87	110.93	525.5
1991	2.9	7.56	5.5	5.32	21617.8	80.5	8478.02	10377.7	119.77	43.66	217.12	605.6
1992	5.4	9.18	4	5.51	26638.1	43.5	13580.77	12537.3	581.24	110.08	194.43	693.2
1993	13.2	10.98	3.5	5.75	34634.4	-122.2	19211.61	15774.6	1114.36	275.15	211.99	835.7
1994	21.7	10.98	5	8.61	4662.3	54	26949.28	20925.8	826.8	337.67	516.2	928.1
1995	14.8	10.98	6	8.35	58260.5	167	39620.83	27082.7	912.82	375.21	735.97	1065.9
1996	6.1	7.47	5	8.32	67800	122.2	34055.32	32322.9	732.76	417.26	50.29	1162.8
1997	0.8	7.47	5.25	8.29	74462.6	404.2	28936.78	35035.6	510.03	452.57	398.9	1309.6
1998	-1.7	4.59	5.47	8.28	79552.8	435.7	24281.39	37093.5	521.02	454.63	449.59	1460.4
1999	-3	5.33	5.33	8.28	82054	292.3	19214.44	39510.2	412.23	403.19	546.75	1518.3
2000	-1.5	6.46	6.46	8.28	89404	241.1	30340.47	43000.25	623.8	407.15	655.74	1457.3

18.2.2　实现方法分析

汇率是在商品交易和货币运动越出国界时产生的，是一国货币价值在国际上的又一表现。因为一国货币汇率受制于经济、政治、军事和心理等因素的影响，这些因素彼此之间既相互联系又相互制约，而且在不同时间，各因素产生作用的强度也会出现交替变化，所以很难准确地找出究竟哪些因素影响着一国货币汇率的变化。

在开放经济中，汇率是一种重要的资源配置价格。汇率的失衡或错估，不仅会破坏经济的外部平衡，而且会给国内宏观经济稳定和经济可持续增长带来一系列不利影响。另外，汇率的变化还能对人们的日常生活和企业的生产销售产生较大的影响。所以，对影响汇率的因素进行分析和探讨，对于指导汇率政策的制定、预测汇率变化趋势、优化投资策略，

以及研究与汇率有关的生产消费等问题都有重要的应用价值。

可以使用探索性数据分析来解决本例问题。

探索性数据分析是 20 世纪后半叶在西方最先兴起的统计分析方法。探索性数据分析可以作为我们认识数据和问题的有力工具，同时也是正式建立统计分析模型之前的铺垫，是科学统计分析的一个重要环节。许多研究者往往一开始就应用经验对模型进行建模分析，而忽略了对统计数据的探索性研究这一过程，这样容易使其分析只停留在问题的表面而不能够深入。

探索性数据分析有如下 3 个特点：

（1）研究直接从原始数据入手。

许多统计方法要先假定数据服从某种分布，然后用适合该种分布的模型进行拟合、分析和预测。但客观实际的数据并不总是满足理论上的分布，因而这些方法具有极大的局限性。探索性数据分析完全从客观数据出发，而不是从某种假定出发，到实际数据中去探索内在的数量规律性。

（2）从实际出发，不以某种理论为依据。

多数的统计模型都是以概率论为理论基础的，并对估计、检验和预测等方法能够给出精确度的度量方法和度量值。探索性数据分析在研究数据的内在特征、数量间的关系和变化时，所用方法会尽可能服从于数据特点和研究目的，并且更重视数据特征的稳健性，而相对忽略对概率论理论和精确度的刻意追求。

（3）分析工具简单直观，更易于普及。

专业统计分析方法运用的数学理论越来越深，这样就使应用的人们越来越害怕统计。探索性数据分析运用简单直观的茎叶图、箱图、残插图和字母值、数据交换、中位数平滑等方法，使具有初等数学知识的人就可以进行分析。

多元回归分析室研究单个因变量和多个自变量之间的相关关系的多元统计方法，比较基础且常用的是多元线性回归分析，许多非线性回归和多项式回归都可以简化为多元线性回归来解决。因而，本例选用多元线性回归分析来分析人民币及其影响因素的相关性。

18.2.3　具体操作步骤

在 SPSS 中进行多元线性回归，可以用“ 分析→回归→线性 ”命令实现，菜单内容如图 18.41 所示。

18.2.3.1　变量的标准化处理

在初始的 12 个自变量里，变量的取值单位有比率、亿元和亿美元等，度量方式不统一，所以有必要先对它们进行标准化处理。

在 SPSS 中，对变量进行标准化处理可以用“ 分析→描述统计→描述 ”命令实现，菜单内容如图 18.42 所示。

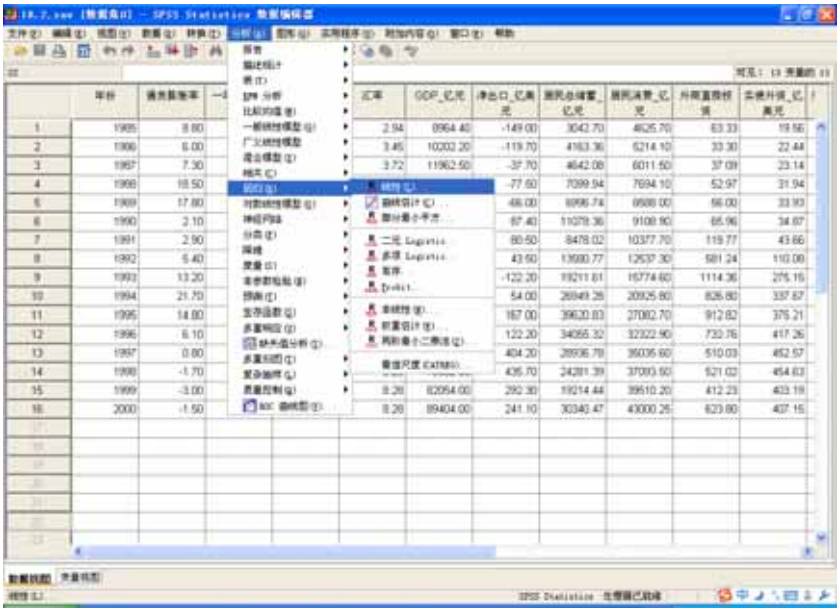


图 18.41　SPSS 多元线性回归模块



图 18.42　SPSS 描述性分析模块

描述性分析的 SPSS 操作如下：

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 描述性分析选入变量及参数设置。

依次单击菜单“ 分析→描述统计→描述 ”，打开“ 描述性 ”对话框，如图 18.43 所示。

在对话框的左侧变量列表框中选择变量“ 通货膨胀率 ”、“ 一年期名义利率 ”、“ 美元利率 ”、“ GDP_亿元 ”、“ 净出口_亿美元 ”、“ 居民总储蓄_亿元 ”、“ 居民消费_亿元 ”、“ 外商直接投资_亿美元 ”、“ 实使外资_亿美元 ”、“ 外汇储备_亿美元 ”和“ 外债规模_亿美元 ”，单击按钮，将其选入到右侧的“ 变量 ”列表框。如图 18.44 所示。



图 18.43 “ 描述性 ”对话框

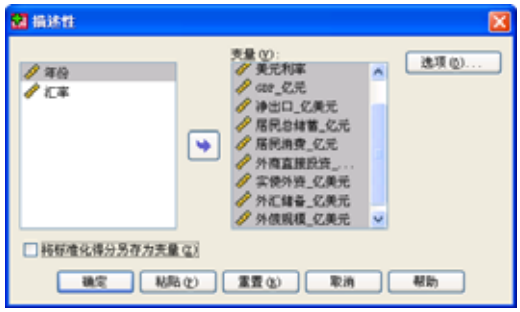


图 18.44　变量选入后的“ 描述性 ”对话框

- 描述性分析：选项对话框

单击图 18.43 中“ 选项 ”按钮，打开“ 描述：选项 ”对话框，如图 18.45 所示。



图 18.45 “ 描述：选项 ”对话框

本例中，依次选择“ 均值 ”；在“ 离散 ”选项组中选择“ 标准差 ”、“ 最小值 ”和“ 最大值 ”；在“ 分布 ”选项组中选择“ 峰度 ”和“ 偏度 ”；在“ 显示顺序 ”选项组中选择“ 变量列表 ”。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 0.
DESCRIPTIVES VARIABLES=通货膨胀率 一年期名义利率 美元利率 GDP_亿元 净出口_亿美元
居民总储蓄_亿元 居民消费_亿元 外商直接投资_亿美元 实使外资_亿美元
    外汇储备_亿美元 外债规模_亿美元
  /STATISTICS=MEAN STDDEV MIN MAX KURTOSIS SKEWNESS.
```

04 描述性分析结果输出。

设置完毕后，单击图 18.43 对话框中的“ 确定 ”按钮，执行描述性分析。

- 描述性统计量

表 18.19　描述性统计量

描述统计量

	N	极小值	极大值	均值	标准差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	统计量	标准误	统计量	标准误
通货膨胀率									
一年期名义利率	16	–3.00	21.70	7.4500	7.72839	.456	.564	–.910	1.091
美元利率	16	4.59	11.34	8.3813	2.08444	–.096	.564	–.886	1 091
GDP_亿元	16	3.50	9.00	6.0944	1.51353	.247	.564	–.466	1.091
净出口_亿美元	16	4662.30	89404.00	38787.5625	30656.51600	.549	.564	–1.513	1.091
居民总储蓄_信元	16	–149.00	435.70	84.7375	183.82747	.606	.564	–.560	1 091
居民消费_亿元	16	3042.70	39620.83	17605.7556	11856.73569	.394	.564	–1.187	1.091
外商直接投资_亿美元	16	4625.70	43000.25	19681.4281	13774.99623	.505	.564	–1.453	1.091
	16	33.30	1114.36	416.4675	363.32503	.412	.564	–1.113	1.091
实使外资_亿美元	16	19.56	454.63	215.1531	186.66141	.110	.564	–2.060	1.091
外汇储备_亿美元	16	20.72	735.97	265.8450	247.44543	.662	.564	–1.018	1.091
外债规模_亿美元	16	158.30	158.30	815.6563	468.79037	.184	.564	–1.406	1.091
有效的 N（列表状态）	16								

表 18.19 为描述性统计量。通过观察每个变量的描述性统计信息，可以了解这个变量的极值情况（极小值和极大值）、取值波动情况（标准差）以及分布情况（峰度和偏度）。从各变量的取值范围来看，相差的数量级很大，所以有必要进行标准化；从峰度、偏度的取值来看（都接近 0），各变量都没有很过分地偏离正态分布。

18.2.3.2　多元线性回归分析

本例的自变量较多，并且它们之间可能存在着共线性问题，故采用逐步回归分析法。

在 SPSS 中，对变量进行多元线性回归分析可以用“ 分析→回归→线性 ”命令实现，菜单内容如图 18.46 所示。

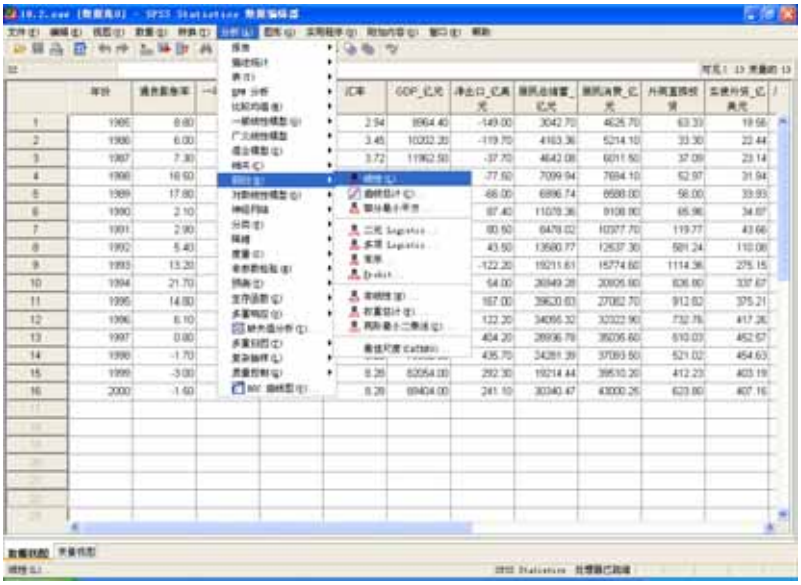


图 18.46　SPSS 多元线性回归分析模块

多元线性回归分析的 SPSS 操作如下：

01 建立数据文件。

将标准化后的数据文件导入到 SPSS 中。

02 描述性分析选入变量及参数设置。

依次单击菜单“ 分析→回归→线性 ”命令，打开“ 线性回归 ”对话框，如图 18.47 所示。

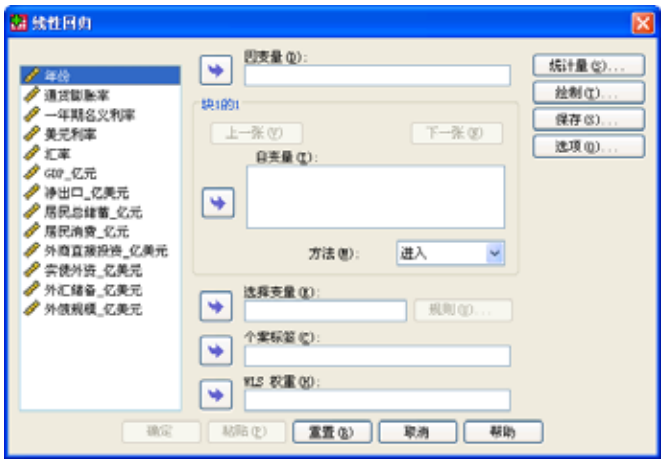


图 18.47　线性回归分析对话框

在对话框的左侧变量列表框中选择变量“ 汇率 ”，单击按钮，将其选入“ 因变量 ”

列表框；按下 ctrl 键，同时选中变量“ 通货膨胀率 ”、“ 一年期名义利率 ”、“ 美元利率 ”、“ GDP_亿元 ”、“ 净出口_亿美元 ”、“ 居民总储蓄_亿元 ”、“ 居民消费_亿元 ”、“ 外商直接投资_亿美元 ”、“ 实使外资_亿美元 ”、“ 外汇储备_亿美元 ”和“ 外债规模_亿美元 ”，单击按钮，将其选入到右侧的“ 自变量 ”列表框，如图 18.48 所示。

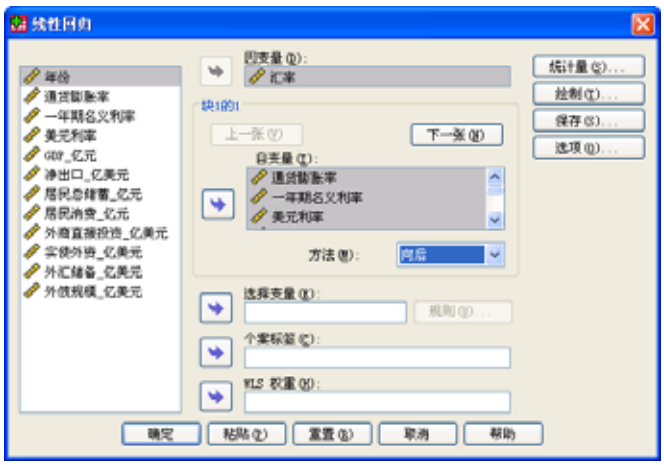


图 18.48　变量选入后的对话框

● 线性回归分析：方法

单击图 18.47 所示对话框“ 方法 ”下拉列表框中单击菜单，选择建立线性回归方程的方法，包括 5 个选项，其内容及具体含义如表 18.20 所示。

表 18.20　方法各选项及其含义

方　法	含　义
进入	强迫引入法，默认选项。全部被选变量一次性进入回归模型，这是最简单的一种回归估计方法
逐步	强迫剔除法。在每一次引入变量时，概率 F 的最小值的变量将引入回归方程。如果已引入回归方程的变量 F 大于设定值，将被剔除出回归方程。当无变量被引入或剔除时，终止回归过程
删除	提出变量。将所有不进入方程模型的被选变量一次性剔除
向后	向后消去法。一次性将所有变量引入方程，并依次进行剔除。首先剔除与因变量最小相关且符合剔除标准的变量，然后进行第二个与因变量最小相关并且符合剔除标准的变量，依次类推。当方程的变量均不满足剔除标准时，终止回归方程
向前	向前引入法。被选变量依次进入回归模型，首先引入与因变量最大相关且符合引入标准的变量，在引入第一个变量后，再引入第二个与因变量最大偏相关并且符合引入标准的变量，依次类推。当无变量符合引入标准时，终止回归过程

经试验发现，向前逐步法只能在最终模型保留一个变量，而向后逐步法只能在最终模型保留多个变量，可见向后逐步回归法更能充分利用本例中的数据。

● 线性回归分析：统计量对话框

单击图 18.47 中的“ 统计量 ”按钮，打开“ 线性回归：统计量 ”对话框。如图 18.49

所示。

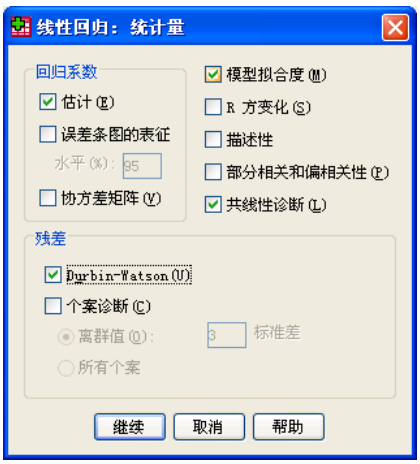


图 18.49 “ 线性回归：统计量 ”对话框

在该对话框中，“ 回归系数 ”选项组控制有关回归系数统计量的显示。其内容及具体含义如表 18.21 所示。

表 18.21　回归系数各选项及其含义

选　　项	含　　义
估计	默认选项。显示回归系数的估计值β、回归系数的标准误差、标准化回归系数、回归系数的β的 t 估计值和双尾显著性水平
误差条图的表征	显示回归系数的β的 95%置信区间
协方差矩阵	显示回归系数的方差—协方差系数矩阵

“ 残差 ”选项组中的选项用于进行有关残差的设置，其内容与具体含义如表 18.22 所示。

表 18.22　残差各选项及其含义

选　　项	含　　义
Durbin-Waston	默认选项。显示残差相关的 Durbin-Waston 检验和残差与预测值的综述统计
个案诊断	选择该项后，在下面的两个单选项中进行选择，确定能够残差分析的个案范围，选项包括“ Outliers outside n standard deviation ”，含义为超过 n 倍标准差以上的个案为奇异值，默认值为 3；“ All cases ”表示显示所有变量的标准化残差、观测值和预测值、残差

其他复选框主要用来控制拟合方程，其选项及含义如表 18.23 所示。

表 18.23　其他选项及其含义

选　　项	含　　义
模型拟合度	默认选项。列出进入或从模型中剔除的变量以及拟合优度统计量、复相关系数（R）、判定系数（R^2）、调整 R^2、估计值的标准误差以及方差分析表

续表

选　项	含　义
R 方变换	增加或删除一个因变量所产生的 R^2 改变量。R^2 改变量越大，表明该变量可能是一个较理想的回归变量
描述性	显示变量的均值、标准差、相关系数矩阵及单尾检验
部分相关和偏相关性	显示零阶相关、偏相关和部分相关系数。方程中至少有两个自变量
共线性诊断	显示包括各变量的容差、方差膨胀因子和共线性的诊断表

本例中，在“ 回归系数 ”选项组中选择“ 估计 ”；在“ 残差 ”选项组中选择“ Durbin-Watson ”；其他复选框中选择“ 模型拟合度 ”和“ 共线性诊断 ”。

- 线性回归分析：图对话框

单击图 18.47 中“ 绘制 ”按钮，打开“ 线性回归：图 ”对话框。如图 18.50 所示。



图 18.50 “ 线性回归：图 ”对话框

该对话框可以分析数据资料的正态性、线性和方差齐性，还可以检测奇异值或异常值等，是研究回归分析中线性规律很有用的工具。

对话框左侧的变量列表框中包含选择绘制轴散点图的坐标轴变量，用户可以从变量列表框中选择变量作为 *X*（横轴变量）轴和 *Y*（纵轴变量）轴；还可以通过单击“ 下一张 ”按钮来重复操作，以获得更多图形。变量列表框其选项及含义如表 18.24 所示。

表 18.24　变量列表框各选项及其含义

选　项	含　义
DEPENDENT	因变量
*ZPRED	标准化预测值
*ZRESID	标准化残差
*DRESID	删除残差
*ADJPRED	调整预测值
*SRESID	Student 残差
*SDRESID	Student 删除残差

在“ 标准化残差图 ”选项区域中选择将要生成的标准化残差图的类型，包括如表 18.25 中的选项。

表 18.25　标准化残差图各选项及其含义

选　　项	含　　义
直方图	标准化残差的直方图，并给出正态曲线
正态概率图	可对标准化残差分布与正态分布进行比较
产生所有部分图	生成每个自变量残差与因变量残差的散点图，最少有两个自变量引入方程才能产生偏残差图

本例中，在左侧变量列表框中选中变量“ *ADJPRED ”，单击按钮，将其选入到“ Y ”列表框中；用同样方法，将变量“ DEPENDENT ”选入到“ X ”列表框；“ 标准化残差图 ”选项组选择“ 正态概率图 ”。

- 线性回归分析：保存对话框

单击图 18.47 中的“ 保存 ”按钮，打开“ 线性回归：保存 ”对话框，如图 18.51 所示。

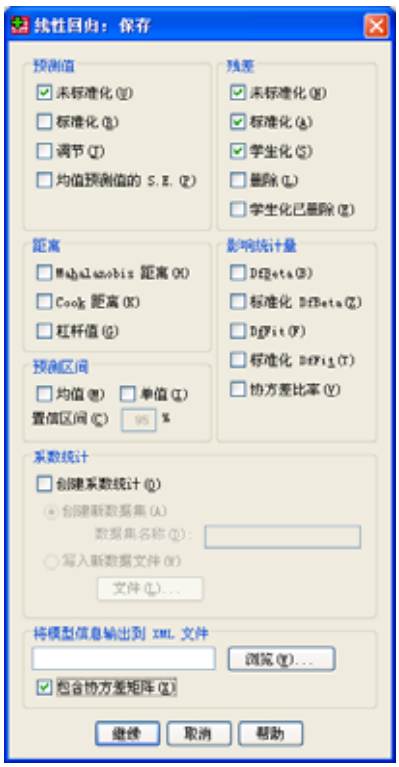


图 18.51 “ 线性回归：保存 ”对话框

利用该对话框，可以选择保存运行过程中生成的部分或全部统计量。对话框中各选项及含义如表 18.26、表 18.27 和表 18.38 所示。

表 18.26　预测值各选项及其含义

选　项	含　义
未标准化	默认选项。非标准化的预测值。模型中因变量的预测值
标准化	标准化的预测值。将每个预测值转换成标准化形式
调节	调整预测值。在回归系数的计算中剔除当前个案时当前个案的预测值
均值预测值的 S.E	预测值均值的标准误差

表 18.27　距离各选项及其含义

选　项	含　义
Mahalanobis 距离	是自自变量个案值与所有个案平均值的距离，Maholanobis 距离大时，表示该个案的一个或多个自变量有奇异值
Cook 距离	当一个个案参与计算回归线系数时，所有个案残差变化的大小，Cook 距离较大时，表示该个案将直接影响回归系数
杠杆值	测量回归拟合点的影响，中心杠杆值的范围是 0 到 *（n-1）/n*

表 18.28　预测区间各选项及其含义

选　项	含　义
均值	平均预测区间，表示平均预测区间与个体预测区间的上下限
单值	个体预测区间，对于单个自变量，因变量平均预测区间的上下限

“ 影响统计量 ”中的选项反映了当剔除某个案的数据以后，回归系数的变化情况，选项如表 18.29 所示。

表 18.29　影响统计量中各选项及其含义

选　项	含　义
DfBeta(s)	DfBeta 值，为删除一个个案后回归系数改变的大小。该值在模型中的每一项均进行计算
标准化 DfBeta(s)	标准化 DfBeta 值，为删除一个个案后回归系数改变的大小。当它大于 2/SQRT(n)时，该点可能为强影响点。该值在模型中的每一项均进行计算
DfFit	DfFit 值，即拟合值之差，删除一个个案后回归系数改变的大小
标准化 DfFit	标准化 DfFit 值，拟合值的标准化差值，删除一个个案后回归系数改变的大小。当它大于 2/SQRT(n)时，该点可能为强影响点。
协方差比率	协方差矩阵的比率，为一个个案不参加回归系数计算时的协方差矩阵与所有个案均参与回归系数计算协方差矩阵的比值。该值若接近于 1，表明该各案对协方差矩阵的影响不大

本例中，在“ 预测值 ”选项组中选择“ 未标准化 ”；在“ 残差 ”选项组中选择“ 未标准化 ”、“ 标准化 ”和“ 学生化 ”；

- 线性回归分析：选项对话框

单击图 18.47 中“ 选项 ”按钮，打开“ 线性回归：选项 ”对话框，如图 18.52 所示。



图 18.52 “ 线性回归：选项 ” 对话框

在“ 线性回归：选项 ” 对话框中，“ 步进方法标准 ” 选项组给出利用逐步回归方法时进入值和删除值的标准，其选项及含义如表 18.30 所示。

表 18.30　步进方法标准中的各选项及其含义

选　项	含　义
使用 *F* 的概率	使用 *F* 显著水平值。当侯选变量中最大 *F* 值的概率 *p* 值小于或等于引入值时，引入相应的变量；在已进入方程的变量中，当最小 *F* 值的概率 *p* 值大于或等于剔除值（默认为 0.10）时，剔除相应的变量。所设定的引入值必须小于剔除值，用户可以设定其他标准，如引入 0.10 剔除 0.11，放宽变量进入方程的标准
使用 *F* 值	使用 *F* 值，含义同上

在默认状态下，回归模型包含“ 在等式中包含常量 ” 选项，若不选此项，则回归模型经过原点。回归方程是否包含常数项，对某些回归结果的解释将不一样。

“ 缺失值 ” 选项组中给出处理缺失值的方式，主要有“ 按列表排除个案 ”、“ 按对排除个案 ” 和“ 使用均值替换 ” 3 种。

03 单击“ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT 汇率
  /METHOD=BACKWARD 通货膨胀率 一年期名义利率 美元利率 GDP_亿元 净出口_亿美元 居民总储蓄_亿元 居民消费_亿元 外商直接投资_亿美元 实使外资_亿美元 外汇储备_亿美元
    外债规模_亿美元
  /SCATTERPLOT=(*ADJPRED ,汇率)
  /RESIDUALS DURBIN NORM(ZRESID)
  /SAVE PRED RESID ZRESID SRESID.
```

04 多元线性回归分析结果输出。

设置完毕后，单击图 18.46 对话框中的“ 确定 ”按钮，执行多元线性回归分析。

- 模型摘要

表 18.31　模型摘要

模型汇总[g]

模　型	R	R 方	调整 R 方	标准估计的误差	Durbin-Watson
1	.999[a]	.999	.995	.15333	
2	.999[b]	.999	.996	.137P8	
3	.999[c]	.999	.997	.12801	
4	.999[d]	.998	.997	.125PO	
5	.999[e]	.998	.996	.13390	
6	.999[f]	.998	.996	.13597	2.943

a．预测变量：（常量），外债规模_亿美元，一年期名义利率，美元得率，外汇储备亿美元，通货膨胀率，居民总储蓄_亿元，将出口_亿美元，GDP_亿元，外商直接投资_亿美元，实使外资_亿美元，居民消费_亿元。

b．预测变量：（常量），外债规模_亿美元，一年期名义利率，美元得率，外汇储备亿美元，居民总储备_亿美元，净出口_亿美元，GDP_亿元，外商直接投资_亿美元，实使外资_亿美元，居民消费_亿元。

c．预测变量：（常量），外债规模_亿元，一年期名义得率，美元利率，外汇储备亿美元，居民总储蓄_亿元，将出口_亿美元，GDP_亿元，外商直接投资_亿美元，实使外资_亿美元。

d．预测变量：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备亿美元，居民总储备_亿元，净出口_亿美元，GDP_亿元，外商直接投资_亿元。

e．预测变量：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备_亿美元，居民量储蓄_亿元，GDP_亿元，外商直接投资_亿美元。

f．预测变量：（常量），外债规模_亿美元，一年期名义利率，美元利率，居民总储蓄_亿美，GDP_亿元，外商直接投资_亿美元。

g．因变量：汇率。

表 18.31 为模型摘要。从表中可以看出，模型摘要给出了逐步回归的各模型的拟合情况，最终模型（模型 6）的 R 值、R 方值和调整 R 方值都达到 0.99 以上，即模型几乎解释了总变异的全部，说明模型的整体拟合效果非常好。

- 方差分析表

表 18.32　方差分析表

Anova[g]

模　型		平 方 和	df	均方	F	Sig.
1	回归	71.881	11	6.535	277.953	.000[a]
	贱差	.094	4	.024		
	总计	71.975	15			

续表

模　型		平 方 和	df	均方	F	Sig.
2	回归	71.881	10	7.188	381.967	.000[b]
	贱差	.094	5	.019		
	总计	71.975	15			
3	回归	71.876	9	7.986	487.379	.000[c]
	贱差	.098	6	.016		
	总计	71.975	15			
4	回归	71.865	8	8.983	573.982	.000[d]
	残差	.110	7	.016		
	总计	71.975	15			
5	回归	71.831	7	10.262	572.309	.000[e]
	贱差	.143	8	.018		
	总计	71.975	15			
6	回归	71.808	6	11.968	647.355	.000[f]
	贱差	.166	9	.018		
	总计	71.975	15			

表 18.31 是多元线性回归分析的方差分析表。从表中可以看出，6 个模型回归的均方分别为 6.535、7.188、7.986、8.983、10.262 和 11.968；剩余的均方分别为 0.024、0.019、0.016、0.016、0.018 和 0.018；*F* 检验统计量的观察值分别为 277.953、381.967、487.379、573.982、572.309 和 647.355，相应的概率ρ值均为 0.000，小于 0.05，可以认为因变量“ 汇率 ”和自变量“ 通货膨胀率 ”、“ 一年期名义利率 ”、“ 美元利率 ”、“ GDP_亿元 ”、“ 净出口_亿美元 ”、“ 居民总储蓄_亿元 ”、“ 居民消费_亿元 ”、“ 外商直接投资_亿美元 ”、“ 实使外资_亿美元 ”、“ 外汇储备_亿美元 ”、“ 外债规模_亿美元 ”存在线性关系。

- 模型的参数估计

表 18.33　模型的参数估计

系数[a]								
模　型		统　计						
		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	（常量）	4.410	.613		7.198	.002		
	通货膨胀率	.000	.020	-.003	-.048	.964	.068	14.747
	一年期名义利率	-.111	.066	-.106	-1.694	.165	.084	11.957
	美元利率	-.190	.063	-.131	-3.018	.039	.173	5.772
	GDP_亿元	-4.010E-5	.000	-.561	-6.937	.002	.050	20.039
	净出口_亿美元	-.001	.001	-.107	-1.053	.352	.032	31.567
	居民总储蓄_亿元	.000	.000	.702	8.646	.001	.050	20.201
	居民消费_亿元	-1.846E-5	.000	-.116	-.426	.692	.004	227.661

续表

系数[a]								
模　型		统　计						
		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	外商直接投资_亿美元	-.002	.001	-.322	-2.803	.049	.025	40.322
	实使外资_亿美元	.001	.002	.107	.657	.547	.012	81.881
	外汇储备_亿美元	.001	.000	.058	1.405	.233	.195	5.129
	外债规模_亿美元	.005	.001	1.027	4.676	.009	.007	147.627
2	（常量）	4.430	.402		11.017	.000		
	一年期名义利率	-.114	.039	-.108	-2.905	.034	.189	5.297
	美元利率	-.191	.053	-.132	-3.609	.015	.196	5.098
	GDP_亿元	-3.999E-5	.000	-.560	-8.416	.000	.059	16.916
	净出口_亿美元	-.001	.001	-.105	-1.282	.256	.039	25.588
	居民总储蓄_亿元	.000	.000	.702	9.667	.000	.050	20.194
	居民消费_亿元	-1.809E-5	.000	-.114	-.474	.656	.005	220.487
	外商直接投资_亿美元	-.002	.001	-.320	-3.257	.023	.027	36.947
	实使外资_亿美元	.001	.001	.103	.886	.416	.020	51.266
	外汇储备_亿美元	.001	.000	.057	1.581	.175	.200	5.004
	外债规模_亿美元	.005	.001	1.025	5.279	.003	.007	144.233
3	（常量）	4.440	.375		11.847	.000		
	一年期名义利率	-.105	.032	-.100	-3.251	.017	.241	4.146
	美元利率	-.193	.049	-.133	-3.937	.008	.198	5.049
	GDP_亿元	-4.128E-5	.000	-.578	-11.326	.000	.088	11.426
	净出口_亿美元	.000	.001	-.076	-1.495	.185	.088	11.313
	居民总储蓄_亿元	.000	.000	.683	12.235	.000	.073	13.684
	外商直接投资_亿美元	-.002	.000	-.289	-4.245	.005	.049	20.353
	实使外资_亿美元	.001	.001	.065	.828	.439	.037	26.923
	外汇储备_亿美元	.000	.000	.050	1.642	.152	.250	3.994
	外债规模_亿美元	.004	.000	.946	10.432	.000	.028	36.088
4	（常量）	4.398	.363		12.118	.000		
	一年期名义利率	-.116	.029	-.110	-4.025	.005	.290	3.452
	美元利率	-.184	.047	-.127	-3.939	.006	.210	4.770
	GDP_亿元	-4.191E-5	.000	-.587	-12.038	.000	.092	10.918
	净出口_亿美元	.000	.001	-.073	-1.471	.185	.089	11.251
	居民总储蓄_亿元	.000	.000	.704	14.484	.000	.092	10.860
	外商直接投资_亿美元	-.002	.000	-.269	-4.321	.003	.056	17.833
	外汇储备_亿美元	.000	.000	.042	1.500	.177	.273	3.666
	外债规模_亿美元	.005	.000	.987	13.406	.000	.040	24.942

续表

系数[a]								
模　型		统　计						
		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
5	（常量）	4.334	.386		11.238	.000		
	一年期名义利率	-.112	.031	-.107	-3.655	.006	.292	3.426
	美元利率	-.157	.046	-.109	-3.414	.009	.246	4.071
	GDP_亿元	-4.134E-5	.000	-.579	-11.163	.000	.093	10.781
	居民总储蓄_亿元	.000	.000	.668	14.901	.000	.124	8.056
	外商直接投资_亿美元	-.001	.000	-.207	-4.227	.003	.104	9.617
	外汇储备_亿美元	.000	.000	.033	1.131	.291	.286	3.495
	外债规模_亿美元	.004	.000	.925	14.383	.000	.060	16.587
6	（常量）	4.192	.370		11.321	.000		
	一年期名义利率	-.112	.031	-.107	-3.591	.006	.292	3.426
	美元利率	-.143	.045	-.099	-3.180	.011	.265	3.779
	GDP_亿元	-4.253E-5	.000	-.595	-11.804	.000	.101	9.901
	居民总储蓄_亿元	.000	.000	.673	14.872	.000	.126	7.968
	外商直接投资_亿美元	-.001	.000	-.200	-4.055	.003	.106	9.467
	外债规模_亿美元	.004	.000	.963	17.315	.000	.083	12.033
a．因变量：汇率								

表 18.33 为模型的参数估计，包含了进入模型的变量，主要描述模型的参数估计值（未标准化的和标准化的），以及每个变量的系数估计值的显著性检验和共线性检验。结果模型（模型 6）中所有变量系数的 *t* 检验 Sig 值都接近或小于 0.01，说明这些系数都显著的不为 0，即这些变量对最终模型的贡献都是显著的。最后一列的共线性统计量，外汇储备的 VIF 值较大（大于 10），故认为它与其他变量间可能存在共线性问题。

- 已排除的变量

表 18.34　已排除的变量

已排除的变量[f]

模　型		BetaIn	t	Sia.	偏 相 关	共线性统计量		
						容差	VIF	最小容差
2	通货膨胀率	–.003[a]	–.048	.964	–.024	.068	14.747	.004
3	通货膨胀率	.002[b]	.031	.977	.014	.070	14.282	.021
	居民消费_信元	–.114[b]	–.474	.656	–.207	.005	220.487	.005
4	通货膨胀率	.025[c]	.556	.598	.221	.124	8.044	.034
	居民消费_信元	.033[c]	.193	.854	.078	.009	115.792	.008
	实使外资_信美元	.065[c]	.828	.439	.320	.037	26.923	.028

续表

模　型		BetaIn	t	Sia.	偏 相 关	共线性统计量		
						容差	VIF	最小容差
5	通货膨胀率	.041[d]	.955	.371	.339	.139	7.213	.056
	居民消费_亿元	.133[d]	.988	.356	.350	.014	73.035	.014
	实使外资_信美元	.056[d]	.663	.528	.243	.037	26.776	.034
	净出口_亿美元	–.073[d]	–1.471	.185	–.486	.089	PP。251	.040
6	通货膨胀率	.030[e]	.689	.510	.237	.144	6.939	.079
	居民消费_亿元	.135[e]	.980	.356	.328	.014	73.031	.014
	实使外资_亿美元	.027[e]	.324	.754	.114	.040	24.709	.040
	净出口_亿美元	–.057[e]	–1.093	.306	–.361	.093	10.728	.044
	外汇储备_亿美元	.033[e]	1.131	.291	.371	.286	3.495	.060

a．模型中的预测变：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备_亿美元，居民总储蓄_亿元，净出口_亿美元，GDP_亿元，外商直接投资_亿美元，实使外资_亿美元，居民消费_亿元。

b．模型中的预测变：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备_亿美元，居民总储蓄_亿元，净出口_亿美元，GDP_亿元，外商直接投资_亿美元，实使外资_亿美元。

c．模型中的预测变：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备_亿美元，居民总储蓄_亿元，净出口_亿美元，GDP_亿元，外商直接投资_亿美元。

d．模型中的预测变：（常量），外债规模_亿美元，一年期名义利率，美元利率，外汇储备_亿美元，居民总储蓄_亿元，GDP_亿元，外商直接投资_亿美元。

e．模型中的预测变：（常量），外债规模_亿美元，一年期名义利率，美元利率，居民总储蓄_亿元，GDP_亿元，外商直接投资_亿美元。

表 18.34 给出的是所有未进入模型的变量检验信息，有 *t* 检验的 Sig 值都大于 0.1 可以看出，这些变量对模型的贡献都不显著，所以它们都不包含在模型里。

- 残差统计量

表 18.35　残差统计量

残差统计量

	极小值	极大值	均值	标准偏差	N
预测值	2.871P	8.5750	6.0850	2.18797	16
标准预测值	–1.469	1.138	.000	1.000	16
预测值的标准误差	.056	.130	.088	.020	16
调整的预测值	2.7934	8.5779	6.0625	2.15329	16
残差	–.18897	.21840	.00000	.10532	16
标准贱差	–1.390	1.606	.000	.775	16
Student 化残差	–1.745	2.217	.042	1.037	16
已删除的残差	–.29794	.41589	.02247	.21342	16
Student 化已删除的残差	–2.023	3.101	.075	1.213	16
Mahal．距离	1.612	12.760	5.625	2.980	16
Cook 的距离	.001	1.149	.190	.307	16
居中杠杆值	.107	.851	.375	.199	16

a．因变量：汇率

表 18.35 为残差统计量。残差统计量是对残差的诊断和分析，给出了预测值、标准预测值、残差和标准残差等的最小值、最大值、均值和标准偏差等统计量。

- 回归标准化残差的标准 P-P 图

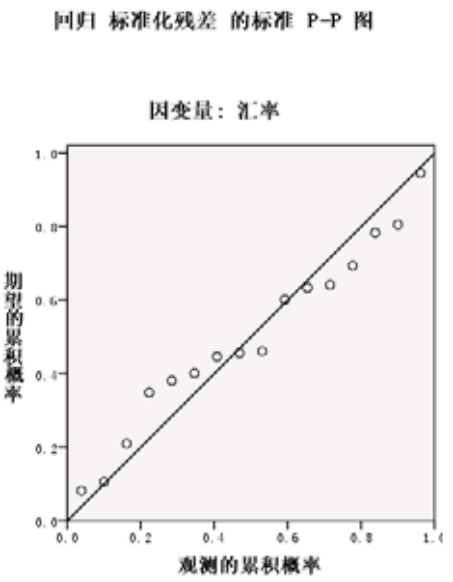


图 18.53　回归标准化残差的标准 P-P 图

图 18.53 是回归标准化残差的标准 P-P 图。该图给出了观察值的残差分布与假设的正态分布比较，如果标准化残差呈正态分布，则标准化的残差散点应分布在直线上或靠近直线。从图 18.54 可以看出，所有残差点都分布在对角线附近，说明残差的正态性假设基本成立。

- 散点图

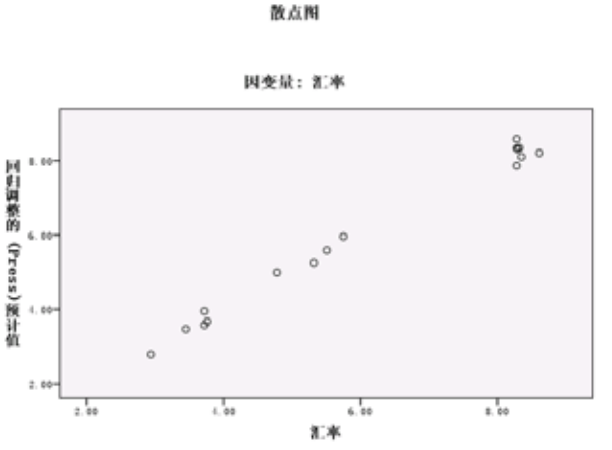


图 18.54　散点图

图 18.53 显示的是因变量与回归标准化预测值的散点图，其中变量“ DEPENDENT ”为 *X* 轴变量，变量“ *ADJPRED ”为 *Y* 轴变量。由图 18.53 可以看出，所有散点都分布在对角线附近，说明预测值和观测值非常接近，预测效果不错。

18.2.3.3　进一步的分析和应用

在前一节建立的多元线性回归模型中，最终模型（模型 6）中的自变量“ 外汇储备_亿美元（标准化）”共线性诊断 VIF 统计量较大，所以本节先把这个变量剔除后重新建立第二个多元线性回归模型，在对结果分析后又剔除新的改进建议，最终得到第三个比较的结果模型。

在图 18.48 所示对话框的“ 自变量 ”列表框中只保留“ 一年期名义利率 ”、“ 美元利率 ”、“ GDP_亿元 ”、“ 居民总储蓄_亿元 ”和“ 外商直接投资_亿美元 ”5 个变量。如图 18.55 所示。

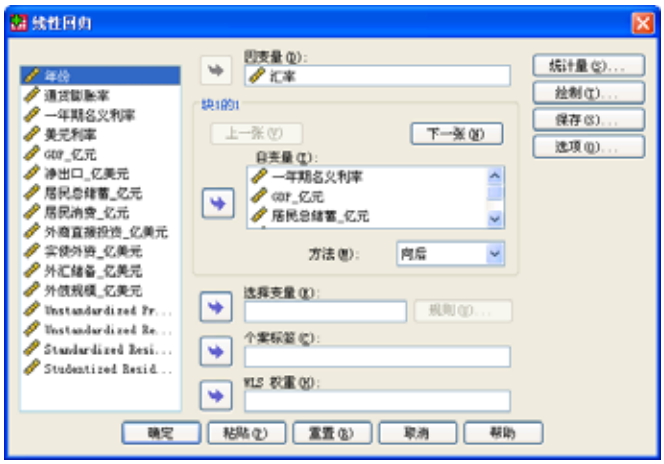


图 18.55　自变量改变后的“ 线性回归 ”对话框

其他设置不变，单击“ 确定 ”按钮，执行多元线性回归分析。

- 模型摘要

表 18.36　模型摘要

模型汇总 [b]

模　型	R	尺方	调整 R 方	标准估计的误差	Durbin-Watson
1	.960[a]	.921	.881	.75561	1.834

a．预测变量：（常量），外商直接投资_亿美元，一年期名义利率，美元利率，GDP_亿元，居民总储蓄_亿元。

b．因变量：汇率

表 18.36 为自变量调整后的模型摘要。从表中可以看出，模型 1 的 R 值大于 90%，R 方值也大于 90%，调整 R 方值大于 85%，说明模型的整体拟合效果很好。

● 方差分析表

表 18.37　方差分析表

Anova[b]

模　　型		平 方 和	df	均 yy	F	Siq.
1	回归	66.265	5	13.253	23.212	.000[a]
	残差	5.709	10	.571		
	总计	71.975	15			

a．预测变量：（常量），外商直接投资_亿美元，一年期名义利率，美元利率，GDP_亿元，居民总储蓄_亿元。

b．因变量：汇率

表 18.37 是自变量调整后的方差分析表。从表中可以看出，模型线性关系成立的 F 检验显著性值 Sig 为 0.000，小于 0.01，说明模型整体的拟合优度不错。

● 模型的参数估计

表 18.38　模型的参数估计

系数 [a]

模　　型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准误差	试用版			容差	VIF
1	（常量）	7.095	1.834		3.868	.003		
	一年期名义利率	–.187	.172	–.178	–1.1088	.302	.298	3.360
	美元利率	–.340	.243	–.235	–1.402	.191	.283	3.540
	GDP_亿元	–1.214E-6	.000	–.017	–.081	.937	.180	5.559
	居民总储蓄_亿元	.000	.000	.956	4.081	.002	.144	6.923
	外商直接投资_亿美元	–.001	.002	–.174	–.637	.538	.1136	9.459

a．因变量：汇率

表 18.38 为自变量调整后模型的参数估计。从表中可以看出，VIF 值均小于 10，说明共线性也有所消除。

18.2.3.4　多元线性回归模型的进一步改进

在 18.2.3.3 节建立的多元线性回归模型中，最终模型（模型 6）包含如下自变量“一年期名义利率”、“美元利率”、“GDP_亿元”、“居民总储蓄_亿元”、“外商直接投资_亿美元”和“外汇储备_亿美元”。其中，“外商直接投资_亿美元”的 VIF 统计量较大，且它和“GDP_亿元”密切相关，考虑到“GDP_亿元”包含的内容更全面，所以建议从模型中剔除“直接投资再次建模。剔除了“外商直接投资_亿美元”的模型结果显示，只有“美元利率”的 t 检验 Sig 值太大，且“外汇储备_亿美元”的 VIF 统计量仍很大，考虑到“美元利率”和“外汇储备_亿美元”之间可能存在的相关关系，下一步建议把 t 检验不显著的

“美元利率”再次剔除。

在图 18.48 所示对话框的“自变量”列表框中只保留“一年期名义利率”、“GDP_亿元”、“居民总储蓄_亿元”和“外汇储备_亿美元”4 个变量，如图 18.56 所示。

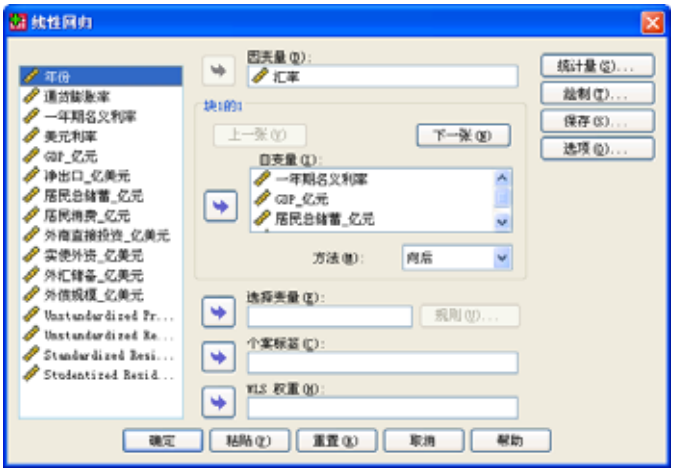


图 18.56　自变量改变后的“线性回归”对话框

其他设置不变，单击“确定”按钮，执行多元线性回归分析。

- 模型摘要

表 18.39　模型摘要

模型汇总[b]

模　型	R	R 方	调整 R 方	标准估计的误差	Durbin-Watson
1	.961[a]	.923	.895	.70996	2.113

a．预测变量：（常量），外商直接投资_亿美元，一年期名义利率，美元利率，GDP_亿元。

b．因变量：汇率

表 18.39 为自变量再次调整后的模型摘要。从表中可以看出，R 方值和调整 R 方值均在 90%左右，说明此模型的拟合优度较表 18.36 要好得多。

- 方差分析表

表 18.40　方差分析表

Anova[b]

模　型		平 方 和	df	均　方	F	Siq.
1	回归	66.430	4	16.608	32.949	.000[a]
	残差	5.544	11	.504		
	总计	71.975	15			

a．预测变量：（常量），外商直接投资_亿美元，一年期名义利率，美元利率，GDP_亿元。

b．因变量：汇率

表 18.40 是自变量再次调整后的方差分析表。从表中可以看出，模型线性关系成立的 *F* 检验显著性值 Sig 为 0.000，小于 0.01，说明模型整体的拟合优度很好。

- 模型的参数估计

表 18.41　模型的参数估计

系数 [a]

模　型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准误差	试用版			容差	VIF
1	（常量）	5.243	1.309		4.006	.002		
	一年期名义利率	–.248	.142	–.236	–7.743	.109	.383	2.608
	GDP_亿元	–6.429E-6	.000	–.090	–.454	.659	.178	5.609
	居民总储蓄_亿元	.000	.000	.802	4.411	.001	.212	4.719
	外汇储备_亿美元	.002	.001	.237	1.794	.100	.400	2 497

a．因变量：汇率

表 18.41 为自变量再次调整后的模型参数估计。从表中可以看出，这些自变量的 VIF 值共线性检验统计量没有特别大的异常值，同时所有变量系数的 t 检验都十分显著，即这些自变量对模型的贡献都是显著的。

从表 18.41 可得改进的汇率预测模型为：

汇率=5.243–0.248*一年期名义得率–6.429E–6*GDP_亿元+0.000*居民总储蓄_亿元+0.002*

这里的自变量都是标准化后的。此方程给出了各因素对汇率的量化影响，其中标准化后的“ GDP_亿元 ”和“ 外汇储备_亿美元 ”对汇率影响最大，且“ GDP_亿元 ”为反向作用，“ 外汇储备_亿美元 ”为正向作用。

采用此改进多元线性回归模型建立的残差 P-P 图和预测趋势图如图 18.57 和图 18.58 所示。

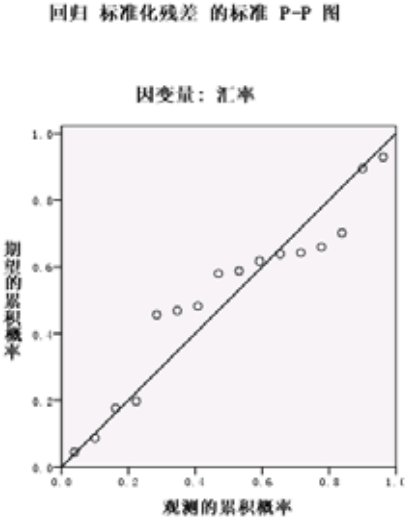


图 18.57　改进回归模型建立的标准化残差 P-P 图

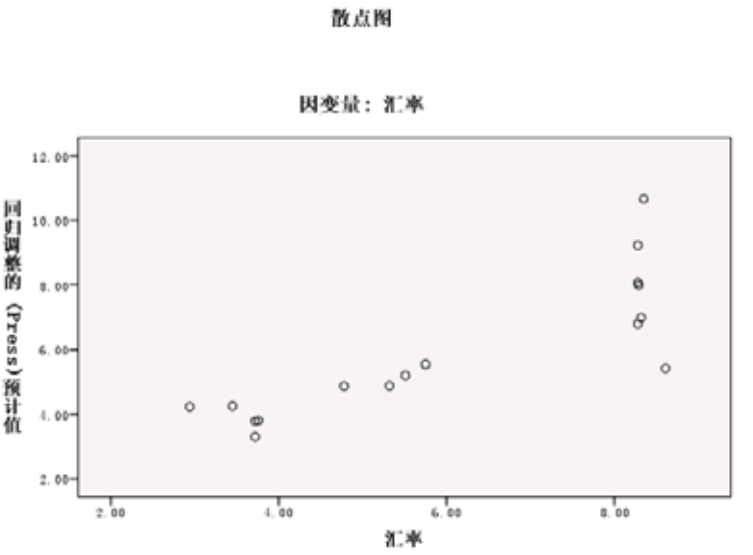


图 18.58　改进回归模型建立的散点图

由图 18.56 和图 18.57 可见，残差的分布基本为正态的，拟合优度较好，精度较高。

18.2.3.5　汇率影响因素的定性分析

许多研究者都多次提到影响汇率的因素是很多的，而且汇率作为经济运行中的一个重要元素，与某些经济指标之间的内在关联在经济学上本身就有合理的理论基础和解释，单纯从数据出发进行的建模甚至有可能违背基本的经济理论，所以从经济学的角度对影响汇率的因素进行分析也是很有必要的。

另外也有研究者认为，对汇率影响因素的分析应从浮动汇率入手，而不建议使用固定汇率，这样更能反映市场自身发展的规律，减少了人为因素。

第19章 SPSS行业应用实例——医学统计

医学是处理健康相关问题的一种科学，以治疗、预防生理和心理疾病、提高人体自身素质为目的。狭义的医学只是疾病的治疗，但也有说法称预防医学为第一医学，临床医学为第二医学，复健医学为第三医学。本章重点结合实例介绍 SPSS 在医学领域的统计分析应用。

19.1 两组不同治疗方法的生存率分析

19.1.1 实例内容说明

用中药+化疗（治疗方法 1，16 例）和单纯化疗（治疗方法 2，10 例）两种疗法治疗白血病患者后，随访存活情况记录如表 19.1 所示，试比较两组的生存率。

表 19.1 两种治疗方式存活情况统计

治疗方法 1		治疗方法 2	
随访月数	是否死亡	随访月数	是否死亡
10	否	2	是
2	是	13	否
12	是	7	是
13	否	11	是
18	否	6	否
6	是	1	否
19	是	11	否
26	否	3	否
9	是	17	否
8	是	7	否

续表

治疗方法 1		治疗方法 2	
随访月数	是否死亡	随访月数	是否死亡
6	是		
43	是		
9	否		
4	否		
31	是		
24	是		

19.1.2 实现方法分析

在科学研究中，可通过随访（Follow Up）来研究事物的变化发展规律，以获取有关生存的信息。随访资料的特点是：

- 需随访一段时间，由于研究时间较长，难免出现研究个体失访等原因而退出研究的现象，形成截尾数据。
- 数据分布类型复杂，一般生存时间数据的分布呈正偏态分布。

正是由于这些缺点，研究分析时不仅需要考虑某种结局（如有效、治愈、死亡等），还需要考虑出现这些结局所经历的时间长短。生存分析通常采用生命表分析、Kaplan-Meier 分析等非参数方法计算与比较单因素生存率；采用 Cox 比例风险回归模型等半参数方法考虑多因素对生存情况的影响。

对于生存资料，首先需要给出各时间点上生存函数的估计值，方法之一是采用生命表法（Lift-Table Method，简称 LT 法）。生命表法适用于区间数据，通过计数落入时间区间内的失效和截尾的观察例数来估计该区间上的死亡概率，然后用该区间及其之前的各区间上的生存概率之积来估计生存率或生存函数。

通过分析发现，本例问题可以使用 SPSS 分析模块中生存函数中的寿命表来解决。

19.1.3 具体操作步骤

在 SPSS 中进行生命表分析，可以利用“分析→生存函数→寿命表”命令实现，菜单内容如图 19.1 所示。

生命表分析的 SPSS 操作如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。在 SPSS 中，随访月数按原数值；已死亡的，为 1，否则为 0；分组方式，治疗方法 1 为 1，治疗方法 2 为 2。

02 选入变量及参数设置。

依次单击菜单“分析→生存函数→寿命表”命令，打开“寿命表”对话框。如图 19.2 所示。

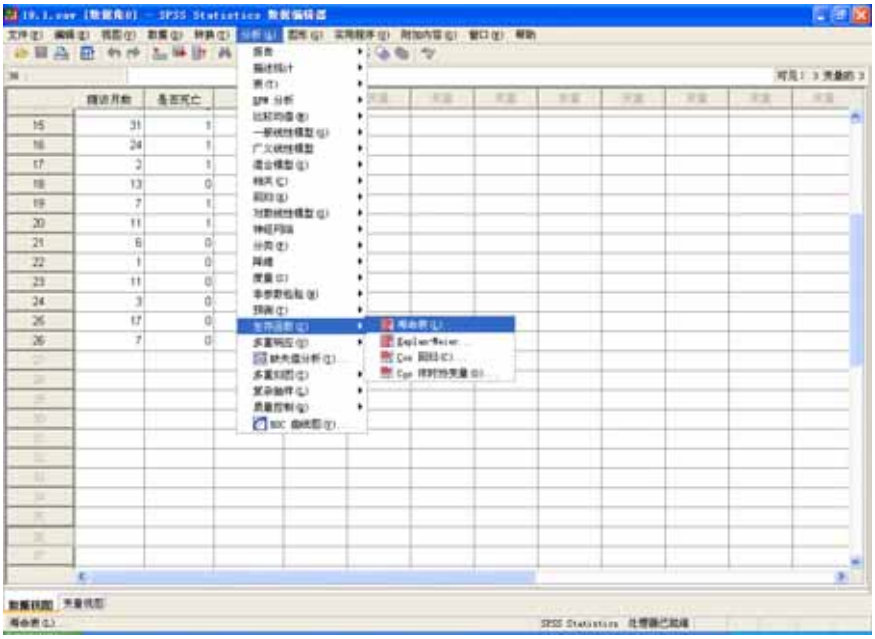


图 19.1　SPSS 寿命表分析模块



图 19.2　“ 寿命表 ” 对话框

在对话框的左侧变量列表框中单击选中变量 “ 随访月数 ”，单击按钮，将其移动到右侧的 “ 时间 ” 列表框。用同样方法将变量 “ 是否死亡 ” 选入到 “ 状态 ” 列表框，将 “ 组别 ” 选入到 “ 因子 ” 列表框。结果如图 19.3 所示。

在图 19.2 所示对话框中，“ 显示时间间隔 ” 栏中定义需要显示生存率的时间，本例要求从 0 个月显示至 48 个月，间隔为 2 个月，故在编辑框 “ 0 到 ” 中输入 48，在编辑框 “ 步长 ” 中输入 2。



图 19.3　选入变量后的“寿命表”对话框

单击图 19.3 中“定义事件”按钮，打开“寿命表：为状态变量定义事件”对话框。在编辑框“单值”中输入 1，表明“是否死亡=1”为发生死亡事件，如图 19.4 所示。

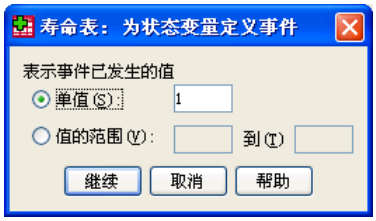


图 19.4　“寿命表：为状态变量定义事件”对话框

单击图 19.2 中“定义范围”按钮，打开“有效表格：定义因子范围”对话框。在“最小值”框中输入 1，“最大值”框中输入 2，表明“是否死亡=1”为发生死亡事件，如图 19.5 所示。

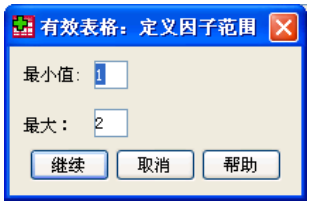


图 19.5　“有效表格：定义因子范围”对话框

- 单变量：模型对话框

单击图 19.2 中“ 选项 ”按钮，打开“ 寿命表：选项 ”对话框，如图 19.6 所示。

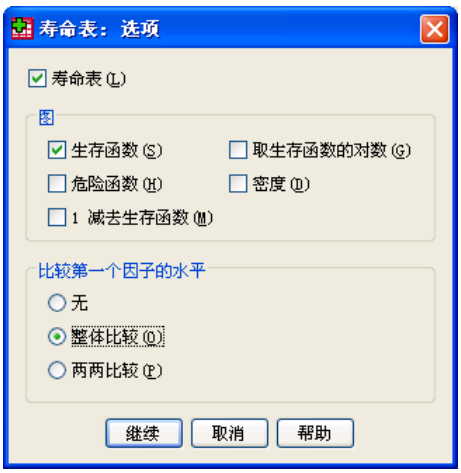


图 19.6 “ 寿命表：选项 ”对话框

该对话框用于选择需要输出的统计量和统计图，包含两个选项组：“ 图 ”和“ 比较第一个因子的水平 ”。

“ 图 ”选项组中各选项及具体含义如表 19.2 所示。

表 19.2　图各选项及其含义

选　　项	含　　义
生存函数	累积生存函数曲线
取生存函数的对数	对数累积生存函数曲线
危险函数	累计风险函数散点图
密度	累计生存函数密度曲线
1 减去生存函数	累计死亡函数曲线

“ 比较第一个因子的水平 ”选项组用于组间比较，有 3 种方式。各选项及其含义如表 19.3 所示。

表 19.3　比较第一个因子水平个选项及其含义

选　　项	含　　义
无	不进行比较
整体比较	从整体上进行比较
两两比较	两个变量之间进行比较

本例中，选择“ 寿命表 ”复选框；在“ 图 ”选项组中选择“ 生存函数 ”；在“ 比较第一个因子的水平 ”选项组中选择“ 整体比较 ”。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集0.
SURVIVAL TABLE=随访月数 BY 组别(1 2)
  /INTERVAL=THRU 48 BY 2
  /STATUS=是否死亡(1)
  /PRINT=TABLE
  /PLOTS (SURVIVAL)=随访月数 BY 组别
  /COMPARE=随访月数 BY 组别.
```

04 结果输出。

设置完毕后，单击“确定”按钮，执行寿命表分析。

- 寿命表

表 19.4　寿命表

年 限 表

一阶控制		期初时间	期初记入数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累计生存比例	期末的累计生存比例的标准误	概率密度	概率密度的标准误	风险率	风险率的标准误
组别	1	0	16	0	16.000	0	.00	1.00	1.00	.00	.000	.000	.00	.00
		2	16	0	16.000	1	.06	.94	.94	.06	.031	.030	.03	.03
		4	15	1	14.500	0	.00	1.00	.94	.06	.000	.000	.00	.00
		6	14	0	14.000	2	.14	.86	.80	.10	.067	.044	.08	.05
		8	12	1	11.500	2	.17	.83	.66	.12	.070	.046	.10	.07
		10	9	1	8.500	0	.00	1.00	.66	.12	.000	.000	.00	.00
		12	8	1	7.500	1	.13	.87	.58	.13	.044	.042	.07	.07
		14	6	0	6.000	0	.00	1.00	.58	.13	.000	.000	.00	.00
		16	6	0	6.000	0	.00	1.00	.58	.13	.000	.000	.00	.00
		18	6	1	5.500	1	.18	.82	.47	.15	.052	.049	.10	.10
		20	4	0	4.000	0	.00	1.00	.47	.15	.000	.000	.00	.00
		22	4	0	4.000	0	.00	1.00	.47	.15	.000	.000	.00	.00
		24	4	0	4.000	1	.25	.75	.35	.15	.059	.054	.14	.14
		26	3	1	2.500	0	.00	1.00	.35	.15	.000	.000	.00	.00
		28	2	0	2.000	0	.00	1.00	.35	.15	.000	.000	.00	.00
		30	2	0	2.000	1	.50	.50	.18	.15	.088	.073	.33	.31
		32	1	0	1.000	0	.00	1.00	.18	.15	.000	.000	.00	.00
		34	1	0	1.000	0	.00	1.00	.18	.15	.000	.000	.00	.00
		36	1	0	1.000	0	.00	1.00	.18	.15	.000	.000	.00	.00
		38	1	0	1.000	0	.00	1.00	.18	.15	.000	.000	.00	.00
		40	1	0	1.000	0	.00	1.00	.18	.15	.000	.000	.00	.00
		42	1	0	1.000	1	1.00	.00	.00	.00	.088	.073	1.00	.00

续表

一阶控制		期初时间	期初记入数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累计生存比例	期末的累计生存比例的标准误	概率密度	概率密度的标准误	风险率	风险率的标准误
	2	0	10	1	9.500	0	.00	1.00	1.00	.00	.000	.000	.00	.00
		2	9	1	8.500	1	.12	.88	.88	.11	.059	.055	.06	.06
		4	7	0	7.000	0	.00	1.00	.88	.11	.000	.000	.00	.00
		6	7	2	6.000	1	.17	.83	.74	.16	.074	.068	.09	.09
		8	4	0	4.000	0	.00	1.00	.74	.16	.000	.000	.00	.00
		10	4	1	3.500	1	.29	.71	.53	.21	.105	.092	.17	.16
		12	2	1	1.500	0	.00	1.00	.53	.21	.000	.000	.00	.00
		14	1	0	1.000	0	.00	1.00	.53	.21	.000	.000	.00	.00
		16	1	1	.500	0	.00	1.00	.53	.21	.000	.000	.00	.00

表 19.4 为计算得到的寿命表。从表中可以看出，按用户指定，从 0 月起，隔 2 个月直至 42 个月（原指定为 0～48 个月，但因 42 个月后，生存概率已为 0，故 42 个月后至 48 个月的生存状况不再显示），分别显示期初时间、期初计入数、期初退出数、历险数、期间终结数、生存比例、期末的累计生存比例、期末的累计生存比例的标准误、概率密度、概率密度的标准误、风险率和风险率的标准误。在本例，用中药+化疗的方式（治疗方式 1）治疗白血病患者，至 8 个月时，死亡率为 17%，生存概率为 83%，生存率为 66%，风险比例为 10%。至 42 个月时，生存概率和生存率均为 0，此时风险比例为 100%。

对照组同类结果的显示，因在 16 个月时生存概率已为 0，故仅从 0 月起，（隔 2 个月）至 16 个月止。分析显示，单纯用化疗，白血病患者的半数生存率约在 16 个月多一点，比中药组少三个月。

- 中位数生存时间

表 19.5　中位数生存时间

中位数生存时间

一阶控制		中位数时间
组别	1	19.44
	2	16.00

表 19.5 为中位数生存时间。通过表可以看出，治疗方式 1 的 50%生存率为 19.44 个月，治疗方式 2 的 50%生存率为 16.00 个月。

- 整体比较

表 19.6　整体比较

整体比较[a]

Wilcoxon(Gehan)统计量	df	Siq.
. 012	P	.912

a．比较精确的。

表 19.6 为“Wilcoxon(Gehan)统计量”检验方法的整体比较结果。从表中可以看出，检验的自由度均为 1，概率 ρ 值（Sig.）为 0.912。

- 生存函数曲线

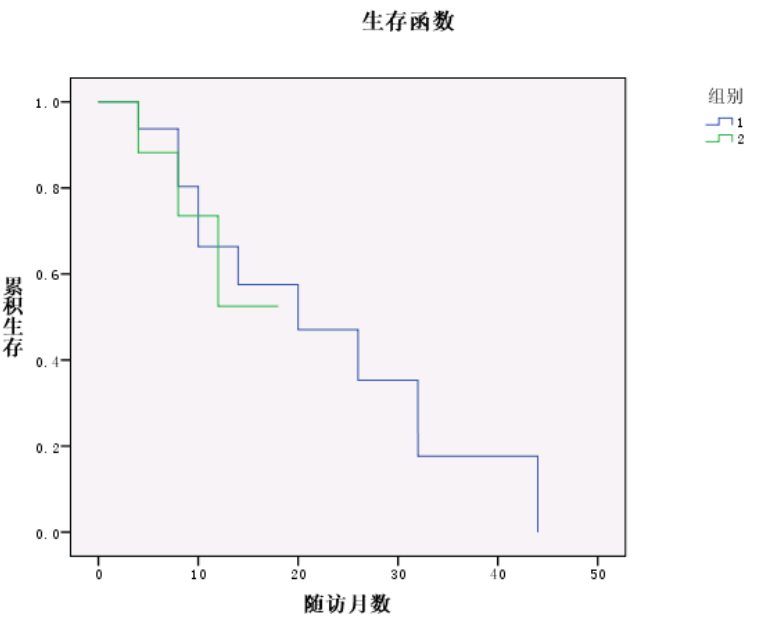


图 19.7　生存函数曲线

图 19.7 为生存函数曲线。从图中可见，治疗方式 2 在 8 个月前一段时点的生存率均较治疗方式 1 略低，而 8～12 个月这一段其生存率又较治疗方式 1 略高，12 个月后再下降。这说明，在治疗中加用中药，对个别患者而言，20 个月后依然有一定的生存率。

19.2　判断不同群体的患病比例有无差异

19.2.1　实例内容说明

用两个班级学生进行两个感冒疫苗的试验，两个班级学生患感冒的情况如表 19.7 所示，问两个班级学生的患病比例有无差别。

表 19.7　两班级学生的患病情况

班　　级	患病人数	未患病人数
1	53	20
2	40	4

19.2.2　实现方法分析

在实际分析中，除了需要对单个变量的数据分布情况进行分析外，还需要掌握多个变量在不同取值情况下的数据分布情况，从而进一步深入分析变量之间的相互影响和关系，这种分析就称为数据交叉分析。

数据交叉分析主要包括两个基本任务：一是根据收集的样本数据，产生二维、多维交叉表；二是在交叉表的基础上，对两两变量间是否存在相关性进行检验。要获得变量之间的相关性，仅靠描述性统计的数据是不够的，还需要借助一些表示变量间相关程度的统计量和一些非参数检验的方法。当所观察的现象同时与两个因素有关时，通过交叉表分析，可以较好地反映出这两个因素之间有无关联性以及两个因素与所观察现象之间的相关关系。

常用的衡量变量间相关程度的统计量是简单相关系数，但在交叉表分析中，由于行列变量往往不是连续变量，不符合计算简单相关系数的前提条件。因此，需要根据变量的性质，选择其他的相关系数，如 Kendall 等级相关系数、Eta 值等。

通过分析发现，本例问题可以使用 SPSS 分析模块中表述统计中的交叉表来解决。

19.2.3　具体操作步骤

在 SPSS 中进行交叉表分析，可以利用“分析→描述统计→交叉表”命令实现，菜单内容如图 19.8 所示。

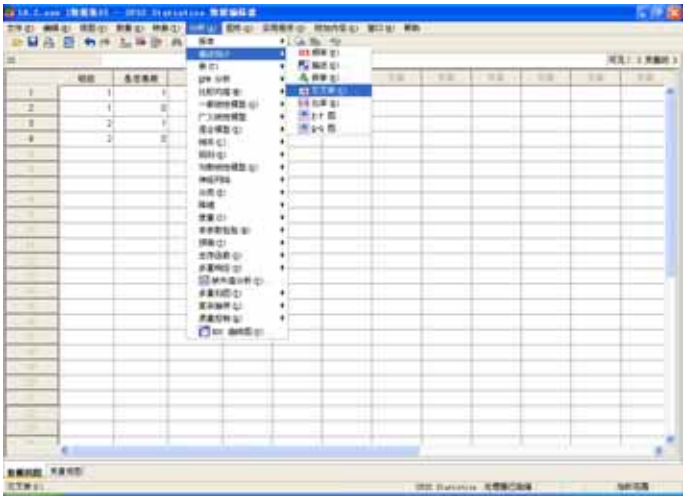


图 19.8　SPSS 交叉表分析模块

交叉表分析的 SPSS 操作如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。在 SPSS 中，变量“是否患病”取值为 1 时表示患病，取值为 0 时表示未患病。

02 变量权重定义。

为了对数据进行进行交叉表分析，首先应该对变量的权重进行定义。在 SPSS 中，可以利用“数据→加权个案”命令实现，菜单内容如图 19.9 所示。

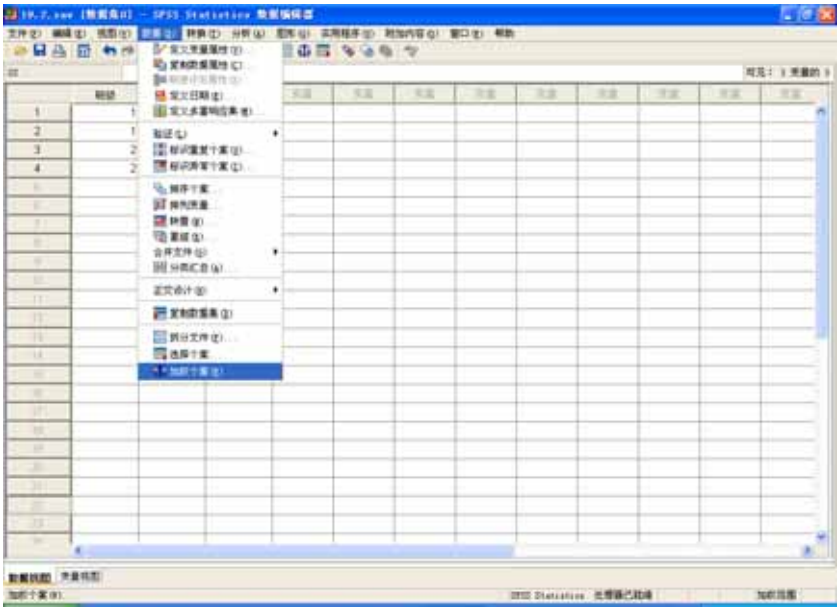


图 19.9　SPSS 加权个案模块

依次单击菜单“数据→加权个案”，打开“加权个案”对话框，如图 19.10 所示。

在图 19.10 对话框中，选择“加权个案”单选按钮，激活“频率变量”列表框，再将左侧变量列表框中的变量“人数”移至右侧“频率变量”列表框中，如图 19.11 所示。

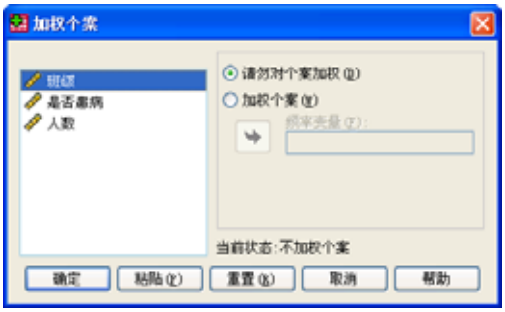


图 19.10　“加权个案”对话框

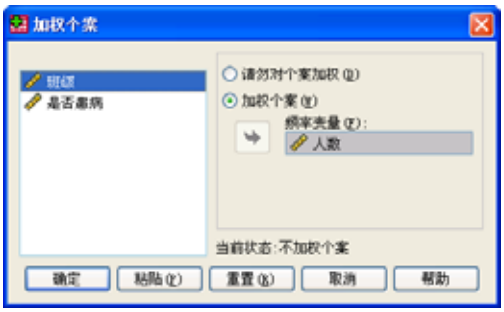


图 19.11　变量选入后的“加权个案”对话框

单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 0.
WEIGHT BY 人数.
```

单击“ 确定 ”按钮，完成对变量的权重的定义。

03 交叉表分析选入变量及参数设置。

依次单击菜单“ 分析→描述统计→交叉表 ”，打开“ 交叉表 ”对话框，如图 19.12 所示。



图 19.12 “ 交叉表 ”对话框

在对话框的左侧变量列表框中选择变量“ 班级 ”，单击按钮，将其选入到右侧的“ 行 ”列表框；用同样方法将变量“ 是否患病 ”选入到“ 列 ”列表框，如图 19.13 所示。



图 19.13　变量选入后的“ 交叉表 ”对话框

● 交叉表分析：精确检验对话框

单击图 19.12 中“ 精确 ”按钮，打开“ 精确检验 ”对话框。如图 19.14 所示。

该对话框提供了三种不同条件的检验方式来检验行列变量的相关性，用户可以选择表 19.8 中所示检验方式。

表 19.8　检验方式各选项及其含义

选　项	含　义
仅渐近法	适用于具有渐近分布的大样本数据，SPSS 默认选择该项
Monte Corlo	此项为精确显著性水平值的无偏估计，无需数据具有渐近分布的假设，是一种非常有效的计算确切显著性水平的方法。在“ 置信水平 ”参数框内输入数据，可以确定置信区间的大小，一般为 90、95 或 99。在“ 样本数 ”参数框中输入数据的样本容量
精确	观察结果概率，同时在下面的“ 每个检验的时间限制为 ”的参数框内，选择进行精确检验的最大时间限度

本例选择默认设置。

● 交叉表分析：统计量对话框

单击图 19.12 中“ 统计量 ”按钮，打开“ 交叉表：统计量 ”对话框，如图 19.15 所示。

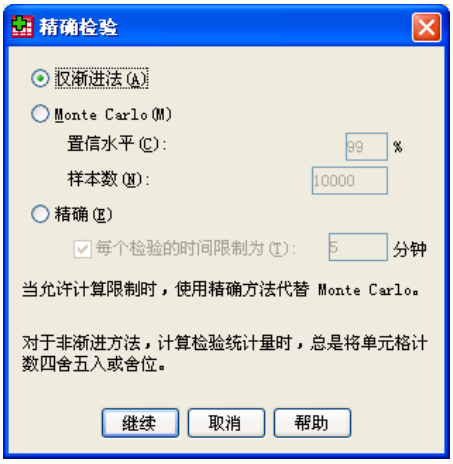


图 19.14 “ 精确检验 ”对话框

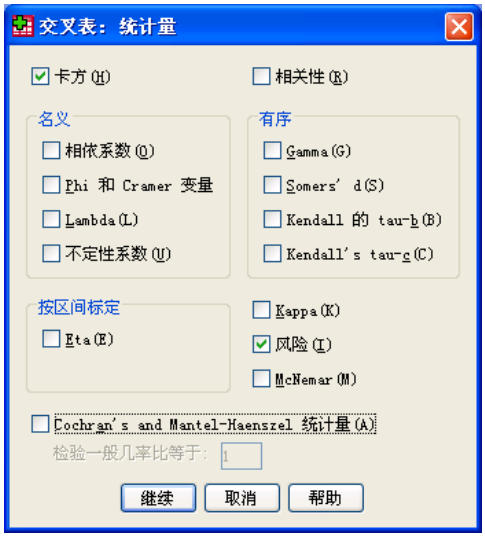


图 19.15 “ 交叉表：统计量 ”对话框

在该对话框中，用户可以选择输出合适的统计检验统计量，对话框中各选项及其含义如下。

卡方和相关性的含义如表 19.9 所示。

表 19.9　卡方和相关性含义

选　项	含　义
卡方	检验列联表行列变量的独立性检验，也被称为 Pearson chi-square 检验、x^2 检验
相关性	输出列联表行列变量的 Pearson 相关系数或 Spearman 相关系数

“名义”选项组适用于名称变量统计量，各选项及其含义如表 19.10 所示。

表 19.10　名义各选项及其含义

选　项	含　义
相依系数	相依系数
Phi and Cramer 变量	Phi and Cramer 变量
Lambda	在自变量预测中用于反映比例缩减误差，其值为 1 时表明自变量预测应变量好，为 0 时表明自变量预测应变量差。
不定性系数	以熵为标准的比例缩减误差，其值接近 1 时表明后一变量的信息在很大程度是来自前一变量，其值接近 0 时表明后一变量的信息与前一变量无关。

“有序”选项组适用于有序变量的统计量，各选项及其含义如表 19.11 所示。

表 19.11　有序各选项及其含义

选　项	含　义
Gamma	两有序变量之间的关联性的对称检验。其数值界于 0~1 之间，所有观察实际数集中在左上角和右下角时，取值为 1，表示两个变量之间有很强的相关，取值为 0 时，表示两个变量之间相互独立
Somers’d	两有序变量之间的关联性的检验，取值范围为[-1，1]
Kendall 的 tau-b	考虑有结的秩或等级变量关联性的非参数检验，相同的观察值选入计算过程中，取值范围为[-1，1]
Kendall 的 tau-c	忽略有结的秩或等级变量关联性的非参数检验，相同的观察值不选入计算过程，取值范围为[-1，1]

“按区间标定”选项组适用于一个名义变量与一个等距变量的相关性检验，各选项及其含义如表 19.12 所示。

表 19.12　按区间标定各选项及其含义

选　项	含　义
Eta	其平方值可认为是应变量受不同因素影响所致方差的比例
Kappa	检验数据内部的一致性，仅适用于具有相同分类值和相同分类数量的变量交叉表
风险	检验事件发生和某因素之间的关联性
McNemar	主要用于检验配对的资料率（相当于配对卡方检验）

本例中，选择“卡方”，“按区间标定”选项组选择“风险”，其他各选项组参数均选择默认。

- 交叉表分析：单元显示对话框

单击图 19.12 中“单元格”按钮，打开“交叉表：单元显示”对话框，如图 19.16 所示。

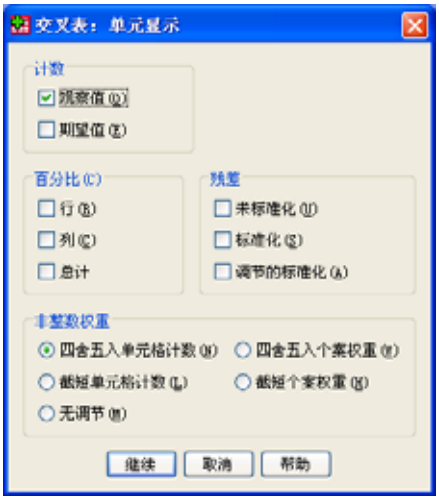


图 19.16 “ 交叉表：单元显示 ” 对话框

在该对话框中，用户可以指定交叉表单元格中的输出内容。具体对话框中各选项的含义如下。

“ 计数 ” 选项组各选项及其含义如表 19.13 所示。

表 19.13　计数各选项及其含义

选　项	含　义
观察值	系统默认选项，表示输出为实际观察值
期望值	表示输出为理论值

“ 百分比 ” 选项组各选项意义如表 19.14 所示。

表 19.14　百分比各选项及其含义

选　项	含　义
行	以行为单元，统计行变量的百分比
列	以列为单元，统计列变量的百分比
总计	行列变量的百分比都进行输出

“ 残差 ” 选项组各选项及其含义如表 19.15 所示。

表 19.15　残差各选项及其含义

选　项	含　义
未标准化	输出非标准化残差，为实际数与理论数的差值
标准化	输出标准化残差，为实际数与理论数的差值除理论数
调节的标准化	输出修正标准化残差，为标准误确定的单元格残差

“ 非整数权重 ” 选项组各选项及其含义如表 19.16 所示。

表 19.16　非整数权重各选项及其含义

选　项	含　义
四舍五入单元格计数	将单元格计数的非整数部分的尾数四舍五入为整数
四舍五入个案权重	将观测量权数的非整数部分的尾数四舍五入为整数
截短单元格计数	将单元格计数的非整数部分的尾数舍去，直接化为整数
截短个案权重	将观测量权数的非整数部分的尾数舍去，直接化为整数
无调节	不对计数数据进行调整

本例“计数”选项组选择“观察值”;“非整数权重”选项组选择“四舍五入单元格计数”;其他选项组选择默认参数。

- 交叉表分析：格式对话框

单击图 19.12 中“格式”按钮，打开“交叉表：表格格式”对话框，如图 19.17 所示。

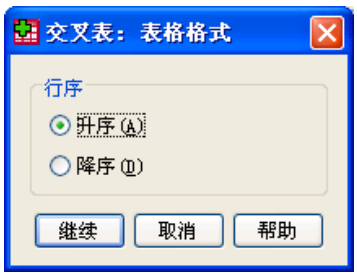


图 19.17 “交叉表：表格格式”对话框

在该对话框中，用户可以指定交叉表的输出排列顺序，具体对话框中各选项的含义如下。

表 19.17　行序各选项及其含义

选　项	含　义
升序	系统默认，以升序显示各变量值。
降序	以降序显示各变量值。

本例选择“升序”。

04 单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
CROSSTABS
  /TABLES=班级 BY 是否患病
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

05 交叉表分析结果输出。

设置完毕后，单击图 19.12 对话框中的“确定”按钮，执行交叉表分析。

● 个案处理摘要

表 19.18　个案处理摘要

个案处理摘要

	案　例					
	有 效 的		缺　失		合　计	
	N	百分比	N	百分比	N	百分比
班级*是否患病	117	100.0%	0	.0%	117	100.0%

表 19.18 为个案处理摘要。表中列出了有效个案数、缺失个案数、总个案数的数量以及所占的百分比。

● 交叉表

表 19.19　交叉表

班级*是否患病交叉表

			是否患病		合　计
			0	1	
班级	1	计数	20	53	73
		期望的计数	15.0	58.0	73.0
		班级中的%	27.4%	72.6%	100.0%
		是否患病中的%	83.3%	57.O%	62.4%
		总数的%	17.1%	45.%	62.4%
		残差	5.0	–5.0	
		标准残差	1.3	–.7	
		调整残差	2.4	–2.4	
	2	计数	4	40	44
		期望的计数	9.0	35.0	44.0
		班级中的%	9.1%	90.9%	100.0%
		是否患病中的%	16.7%	43.0%	37.6%
		总数的%	3.4%	34.2%	37.6%
		残差	–5.0	5.0	
		标准残差	–1.7	.8	
		调整残差	–2.4	2.4	
合计		计数	24	93	117
		期望的计数	24.0	93.0	117.0
		班级中的%	20.5%	79.5%	100.0%
		是否患病中的%	100.0%	100.0%	100.0%
		总数的%	20.5%	79.5%	100.0%

表 19.19 为变量“ 班级 ”和变量“ 是否患病 ”的交叉表。表中给出了两个班级是否患病的人数、期望的计数、占班级总人数的百分比，以及残差、标准残差和调整残差等。

- 风险估计

表 19.20　风险估计

风险估计

	值	95%置信区间	
		下　限	上　限
班级(1/2)的几率比	3.774	1.196	11.970
用于 cohort 是否患病=0	3.014	1.102	8.243
用于 cohort 是否患病=1	799	.674	.946
有效案例中的 N	117		

表 19.20 为风险估计结果。从表中可以看出，在 95%的置信度下，未患病（是否患病=0）的置信区间为（1.102 8.243），患病（是否患病=1）的置信区间为（0.674 0.946）。可以认为这两个班级的患病比例并无明显差异。

- 频数分布条形图

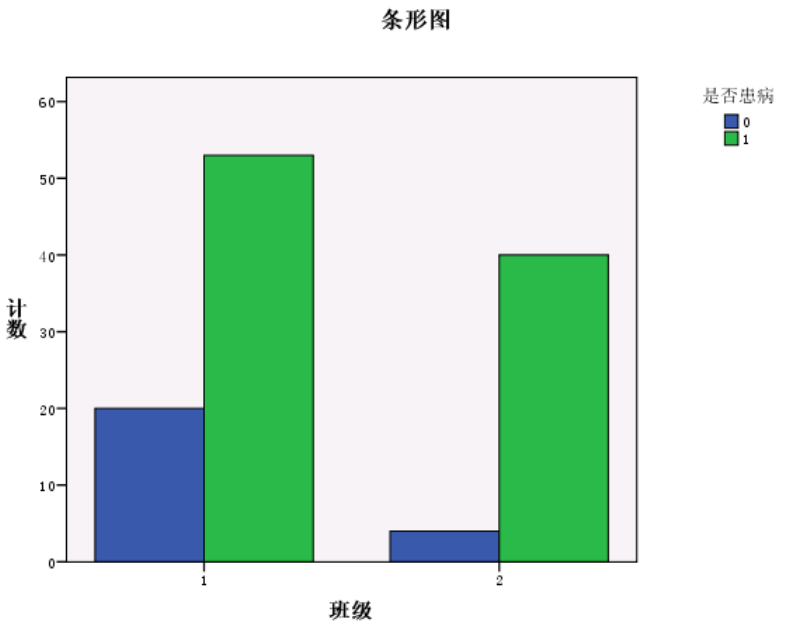


图 19.18　频数分布条形图

图 19.18 为两个班级是否患病人数的频数分布条形图。从图中可以看出，患病人数要高于未患病人数，并且班级 2 的人数要少于班级 1 的人数。

第20章 SPSS行业应用实例——生物学统计

生物学是自然科学的一个门类，它研究生物的结构、功能、发生和发展的规律，以及生物与周围环境的关系。根据研究对象的不同可以分为动物学、植物学、微生物学、古生物学等；根据研究内容的不同，可以分为分类学、解剖学、生理学、细胞学、分子生物学、遗传学、进化生物学、生态学等。生物学作为一门基础科学，传统上一直是农学和医学的基础，涉及种植业、畜牧业、渔业、医疗、制药、卫生等方面。本章重点介绍 SPSS 在医学领域的应用。

20.1 判断某种元素的缺乏是否对另一种元素含量有影响

20.1.1 实例内容说明

某单位研究饲料中缺乏维生素 E 与肝中维生素 A 含量的关系，将大白鼠按性别、体重等配为 8 对，每对中两只大白鼠分别喂给正常饲料和维生素 E 缺乏饲料，一段时间后，测定其肝中维生素 A 含量（μmol/L）如表 20.1 所示，试问饲料中缺乏维生素 E 对鼠肝中维生素 A 含量有无影响。

表 20.1 维生素 A 含量

大白鼠对别	肝中维生素 A 含量（μmol/L）	
	正常饲料组	维生素 E 缺乏饲料组
1	37.2	25.7
2	20.9	25.1
3	31.4	18.8

续表

大白鼠对别	肝中维生素 A 含量（μmol/L）	
	正常饲料组	维生素 E 缺乏饲料组
4	41.4	33.5
5	39.8	34.0
6	39.3	28.3
7	36.1	26.2
8	31.9	18.3

20.1.2　实现方法分析

在实际问题的研究中，通常对研究对象的全部特征或部分特征了解得不是很清楚，如果将研究对象看成一个总体，则可以从总体中抽取样本数据，利用样本数据和已知的部分信息对总体的未知特征进行统计推断，这里的未知特征可以是总体的分布，也可以是总体的某些参数，这就是统计学中的统计推断问题，也称为参数检验。参数检验包含两类有密切关系的问题：参数估计和假设检验。

SPSS 中参数检验主要包括单样本 T 检验、独立样本 T 检验和配对样本 T 检验。

（1）单样本 T 检验

单样本 T 检验是检验某个变量的总体均值和某指定值之间是否存在显著差异，统计的前提是样本总体服从正态分布，也就是说单样本本身无法比较，进行的是其均数与已知总体均数间的比较。

（2）独立样本 T 检验

独立样本的 T 检验用于检验两个独立样本是否来自于具有相同均值的总体，也就是检验两个独立正态总体的均值是否相等。

（3）配对样本 T 检验

配对样本 T 检验用于检验两个相关样本是否来自于具有相同均值的总体，即通过对两个配对样本的 T 检验，推断两个总体的均值是否存在显著差异。

通过分析发现，本例问题可以用 SPSS 分析模块中比较均值中的配对样本 T 检验来解决。

配对的概念指的是两个样本的各样本值之间存在着对应关系。例如，分析一组病人治疗前和治疗后的体重变化，显然，同一个病人对应治疗前后两种不同的体重，同一个病人的两种体重就是一个配对，若干个病人治疗前后两组体重构成了配对样本。从上面可以看出，配对样本的两个样本值之间的配对是一一对应的，并且两个样本的容量相同。配对样本的 T 检验是通过求出每个观测量值之差，所有样本值的观测值之差形成了一个新的单样本，显然，如果两个样本的均值没有显著差异，则样本值之差的均值应该接近零，这实际上转换成为了一个单样本的 T 检验。所以，配对样本的 T 检验就是检验差值的来自总体均值是否为零，这里要求差值来自总体的服从正态分布。

20.1.3 具体操作步骤

在 SPSS 中进行配对样本 T 检验分析，可以利用“ 分析→比较均值→配对样本 T 检验 ”命令实现，菜单内容如图 20.1 所示。

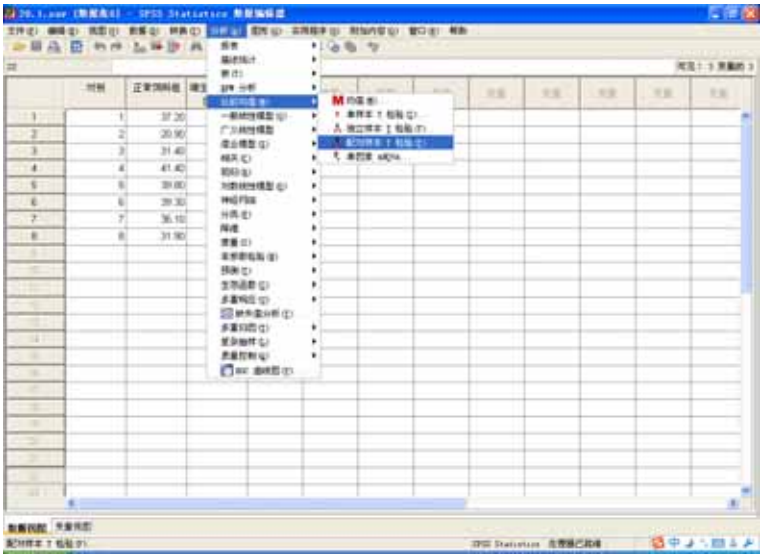


图 20.1 SPSS 配对样本 T 检验模块

配对样本 T 检验的 SPSS 操作如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“ 分析→比较均值→配对样本 T 检验 ”，打开“ 配对样本 T 检验 ”对话框，如图 20.2 所示。

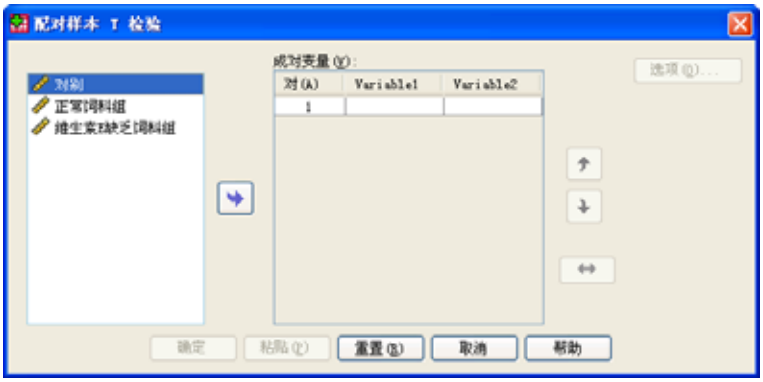


图 20.2 “ 配对样本 T 检验 ”对话框

在对话框的左侧变量列表框中单击选中变量“ 正常饲料组 ”，单击按钮，将其移动到右侧的“ 成对变量 ”列表组的“ Variable1 ”框中。用同样方法将变量“ 维生素 E 缺乏饲料组 ”选入到“ Variable2 ”框。结果如图 20.3 所示。



图 20.3　选入变量后的“ 配对样本 T 检验 ”对话框

- 配对样本 T 检验：选项对话框

单击图 20.2 中“ 选项 ”按钮，打开“ 配对样本 T 检验：选项 ”对话框。如图 20.4 所示。

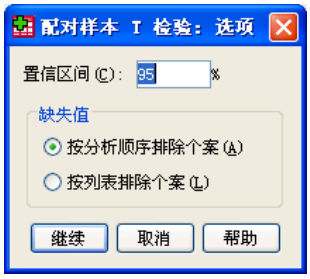


图 20.4　“ 配对样本 T 检验：选项 ”对话框

该对话框用于指定置信区间和缺失值的处理方法。

在“ 信赖区间 ”文本框中输入置信水平，默认置信水平为 95%，也可以输入 1-99 之间的任意整数。在“ 缺失值 ”选项组中指定缺失值的处理方式，有两个单选项，分别为：“ 按分析顺序排除个案 ”和“ 按列表排除个案 ”。

03 单击“ 粘贴 ”按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 0.
T-TEST PAIRS=正常饲料组 WITH 维生素 E 缺乏饲料组 (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

04 结果输出。

设置完毕后，单击“ 确定 ”按钮，执行配对样本 T 检验。

- 成对样本统计量

表 20.2　成对样本统计量

成对样本统计量

	均　值	N	标 准 差	均值的标准误
对 1　正常饲料组	34.7500	8	6.64852	2.35061
维生素 E 缺乏饲料组	26.2375	8	5.82064	2.0579P

表 20.2 为成对样本统计量。表中给出了各组观测量的均值、个案数、标准差和均值的标准误。正常饲料组和维生素 E 缺乏饲料组的均值分别为 34.7500 和 26.2375，标准差为 6.64852 和 5.82064，均值的标准误为 2.35061 和 2.05791

- 成对样本相关系数

表 20.3　成对样本相关系数

成对样本相关系数

	N	相关系数	Siq
对 1　正常饲料组&维生素 E 缺乏饲料组	8	.586	.127

表 20.3 为成对样本相关系数。从表中可以看出，本例共有 8 对观察值，正常饲料组和维生素 E 缺乏饲料组的相关系数为 0.586，相关系数的显著性检验（Sig）为 0.127。

- 成对样本检验

表 20.4　成对样本检验

成对样本检验

	成对差分					t	df	Siq（双侧）
	均值	标准差	均值的标准误	差分的 95%置信区间				
				下限	上限			
对 1　正常饲饲料组–维生索 E 缺乏饲料组	8.51250	5.71925	2.02206	3.73109	1.329391	4.210	7	.004

表 20.4 为成对样本检验。从表中可以看出，正常饲料组和维生素 E 缺乏饲料组两两相减的差值均值、标准差和均值的标准误差分别为 8.51250、5.71925 和 2.02206，差分的 95%置信区间分别为（3.73109 13.29391）。配对检验结果的 T 检验值为 4.210，自由度为 7，双侧显著性检验（Sig）为 0.004。

通过以上分析可知，这两组的差别具有高度显著性意义，即饲料中缺乏维生素 E 对鼠肝中维生素 A 的含量确实有影响。

20.2　根据动物身体属性数据分析个体之间的相似性

20.2.1　实例内容说明

某动物一次产下 3 个幼仔，分别对 3 个幼仔的长度、体重、四肢总长、头重进行测量，试就这几个测量，分析 3 个幼仔的相似性，数据如表 20.5 所示。

表 20.5　幼仔身体属性数据

序　　号	长度（毫米）	体重（克）	四肢总长（毫米）	头重（克）
1	50	215	100	11
2	51	220	110	12
3	52	220	112	12

20.2.2　实现方法分析

描述变量之间线性相关程度的强弱，并用适当的统计指标表示出来的过程为相关分析。可根据研究的目的不同，或变量的类型不同，采用不同的相关分析方法。

SPSS 中相关分析主要有双变量相关分析、偏相关分析和距离相关分析。

（1）双变量相关分析

双变量相关分析是指通过计算变量间两两相关的相关系数，对两个或两个以上变量之间两两相关的程度进行分析。根据所研究的变量类型不同，又可以分为二元定距变量的相关分析和二元定序变量的相关分析。二元定距变量的相关分析是指通过计算定距变量间两两相关的相关系数，对两个或两个以上定距变量之间两两相关的程度进行分析；而定序变量又称为有序（ordinal）变量、顺序变量，它取值的大小能够表示观测对象的某种顺序关系（等级、方位或大小等），也是基于“ 质 ”因素的变量。

（2）偏相关分析

偏相关分析是指当两个变量同时与第三个变量相关时，将第三个变量的影响剔除，只分析另外两个变量之间相关程度的过程。

（3）距离相关分析

距离相关分析是对观测量之间或变量之间相似或不相似的程度的一种测量。距离相关分析可用于同一变量内部各个取值间，以考察其相互接近程度；也可用于变量间，以考察预测值对实际值的拟合优度。

通过分析发现，本例问题可以用 SPSS 分析模块中相关分析中的距离相关分析来解决。

20.2.3　具体操作步骤

在 SPSS 中进行距离相关分析，可以利用“ 分析→相关→距离 ”命令实现，菜单内容

如图 20.5 所示。

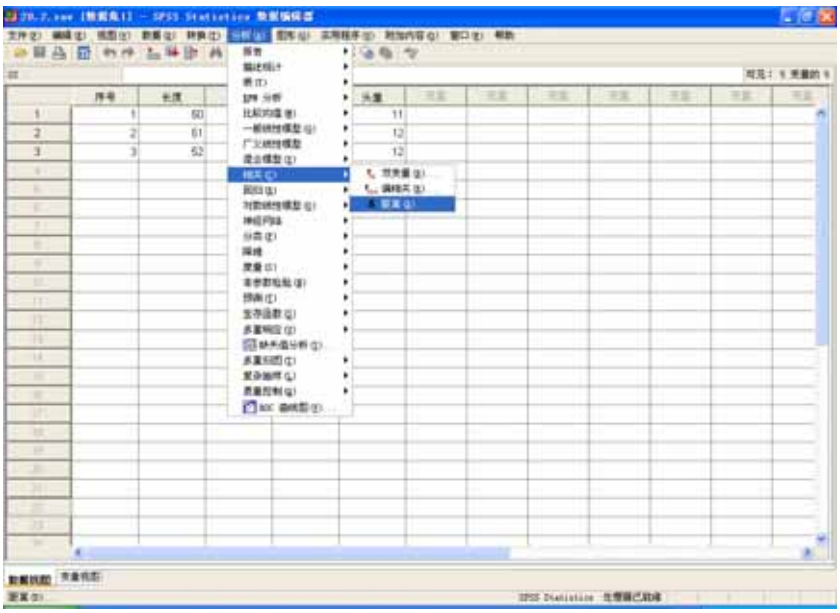


图 20.5　SPSS 距离相关分析模块

距离相关分析的 SPSS 操作如下。

01 建立数据文件。

将数据文件导入到 SPSS 中。

02 选入变量及参数设置。

依次单击菜单“分析→相关→距离”，打开“距离”对话框，如图 20.6 所示。



图 20.6 “距离”相关分析对话框

按下 Ctrl 键，在对话框的左侧变量列表框中单击选中变量“ 长度 ”、“ 体重 ”、“ 四肢总长 ”和“ 头重 ”，单击按钮，将其移到右侧的“ 变量 ”列表框中，如图 20.7 所示。



图 20.7　选入变量后的“ 距离 ”对话框

- 距离：非相似性度量对话框

在图 20.2“ 度量标准 ”选项组中选择“ 不相似性 ”，单击“ 度量 ”按钮，打开“ 距离：非相似性度量 ”对话框，如图 20.8 所示。

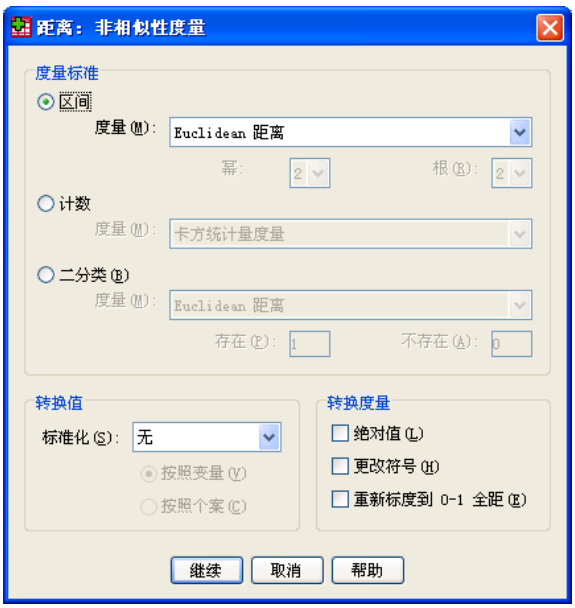


图 20.8　“ 距离：非相似性度量 ”对话框

该对话框用于指定非相似性度量的处理方法，用户可以选择“度量标准”、“转换值”和“转换度量”等。

- 距离：相似性度量对话框

在图 20.2“度量标准”选项组中选择“相似性”，单击“度量”按钮，打开“距离：相似性度量”对话框。如图 20.9 所示。



图 20.9 “距离：相似性度量”对话框

该对话框用于指定相似性度量的处理方法。用户可以选择“度量标准”、“转换值”和“转换度量”等。

03 相似性度量结果。

在图 20.6“度量方式”选项组中选择“相似性”。

单击“粘贴”按钮，得到的 Syntax 命令语句程序如下：

```
PROXIMITIES 长度 体重 四肢总长 头重
  /VIEW=CASE
  /MEASURE=CORRELATION
  /STANDARDIZE= NONE.
```

设置完毕后，单击“确定”按钮，执行距离相关分析。

- 个案处理摘要

表 20.6 个案处理摘要

案例处理摘要

案　例					
有　效		缺　失		合　计	
N	百分比	N	百分比	N	百分比
3	100.0%	0	.0%	3	100.0%

表 20.6 为相似性度量情况下的个案处理摘要。表中给出了各组观测量的有效数、缺失数和百分比。

- 近似矩阵

表 20.7　近似矩阵

近似矩阵

	值向量间的相关性		
	1	2	3
1	1.000	.999	.999
2	.999	1.000	1.000
3	.999	1.000	1.000

表 20.7 为相似性度量情况下个案之间的近似矩阵。由表 20.7 可以看出，不同个体之间的相关系数均在 0.999 以上，这说明三个幼仔个体之间具有较强的相似性，其中个体 2 和个体 3 相似性最大。

04 非相似性度量结果。

将图 20.6 “ 度量方式 ” 改为 “ 不相似性 ”，其他参数设置不变，如图 20.10 所示。



图 20.10　度量方式改变后的 “ 距离 ” 对话框

单击 “ 粘贴 ” 按钮，得到的 Syntax 命令语句程序如下：

```
DATASET ACTIVATE 数据集 1.
PROXIMITIES 长度 体重 四肢总长 头重
  /VIEW=CASE
  /MEASURE=EUCLID
  /STANDARDIZE=NONE.
```

设置完毕后，单击 “ 确定 ” 按钮，执行距离相关分析。

- 个案处理摘要

表 20.8　个案处理摘要

案例处理摘要

案　例					
有　效		缺　失		合　计	
N	百分比	N	百分比	N	百分比
3	100.0%	0	.0%	3	100.O%

表 20.8 为非相似性度量情况下的个案处理摘要。表中给出了各组观测量的有效数、缺失数和百分比。

- 非相似性矩阵

表 20.9　非相似性矩阵

近似矩阵

	Euclidean 距离		
	1	2	3
1	.000	11.269	13.191
2	11.269	.000	2.236
3	13.191	2.236	.000

这是一个不相似性矩阵

表 20.9 为非相似性度量情况下个案之间的非相似性矩阵。由表可以看出，相同个体之间的非相似性为 0.000，个体 1 和个体 2 之间的非相似性为 11.269，个体 2 和个体 3 之间的非相似性为 2.236，个体 1 和个体 3 之间的非相似性为 13.191。这说明个体 2 和个体 3 之间的相似性最大，个体 1 和个体 2 次之，与前述分析一致。